

INTERNATIONAL SEARCH REPORT

I. International application No.

PCT/US 99/ 18424

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 1
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out specifically:
see FURTHER INFORMATION sheet PCT/ISA/210

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of Item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International Application No. PCT/US 99 18424

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box I.2

Claims Nos.: 1

Present claim 1 relate to an extremely large number of possible compounds, i.e. monooxygenases. Support within the meaning of Article 6 PCT and/or disclosure within the meaning of Article 5 PCT is to be found, however, for only a very small proportion of the monooxygenases claimed. In the present case, the claims so lack support, and the application so lacks disclosure, that a meaningful search over the whole of the claimed scope is impossible. Consequently, the search has been carried out for those parts of the claims which appear to be supported and disclosed, namely those parts relating to the monooxygenase references incorporated in the description i.e. P450BM-P (CYP102) described in pages 37-38, and 43 (for the construction of chimeric P450s); styrene monooxygenase *Pseudomonas* sp. strain VLB120 (stdSc, stdR, stdA, stdB, stdC and stdD genes), described in page 47 and 52 and corresponding to the GenBank accession number AF031161 (for the epoxidation of olefins and degradation of methyl-substituted aromatic compounds); P450 CYP2B subfamily described in page 50 (for omega-hydroxylation of fatty acids and detoxification activity); P450cam (CYP2C9) described in page 54 (for the dehydrogenation reactions); P4503A described in page 56 (for the obtention of cyclosporin); P450sca described in pages 56-57 (for the obtention of pravastin); SuaC CYP105A1 and SubC CYP105B1 described in pages 58 (for herbicide resistance and bioremediation) and *Pseudomonas putida* OUS82 as described in page 86, corresponding to GenBank accession number AB004059 (monooxygenases acting as dioxygenases, for alkyl group monooxygenation).

The applicant's attention is drawn to the fact that claims, or parts of claims, relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/US 99/18424

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9720078 A	05-06-1997	US 5811238 A	22-09-1998
		AU 1087397 A	19-06-1997
		AU 2542697 A	17-10-1997
		CA 2239099 A	05-06-1997
		EP 0876509 A	11-11-1998
		EP 0906418 A	07-04-1999
		EP 0911396 A	28-04-1999
		WO 9735966 A	02-10-1997
		US 5837458 A	17-11-1998
WO 9813485 A	02-04-1998	AU 4503797 A	17-04-1998
		AU 4597197 A	17-04-1998
		WO 9813487 A	02-04-1998
WO 9534679 A	21-12-1995	AU 2860295 A	05-01-1996
		GB 2303853 A,B	05-03-1997
		US 5891633 A	06-04-1999
JP 5049474 A	02-03-1993	JP 1888278 C	07-12-1994
		JP 6030584 B	27-04-1994
		JP 62104583 A	15-05-1987
		JP 2099129 C	22-10-1996
		JP 8013270 B	14-02-1996



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12N 15/82, 15/52, 9/10, 1/21, A01H 5/00, C12Q 1/25	A2	(11) International Publication Number: WO 99/67402 (43) International Publication Date: 29 December 1999 (29.12.99)
(21) International Application Number: PCT/EP99/04309 (22) International Filing Date: 22 June 1999 (22.06.99) (30) Priority Data: 09/103,895 24 June 1998 (24.06.98) US (71) Applicant (for all designated States except AT US): NOVARTIS AG [CH/CH]; Schwarzwaldallee 215, CH-4058 Basel (CH). (71) Applicant (for AT only): NOVARTIS-ERFINDUNGEN VERWALTUNGSGESELLSCHAFT M.B.H. [AT/AT]; Brunner Strasse 59, A-1230 Vienna (AT). (72) Inventors; and (75) Inventors/Applicants (for US only): WARD, Eric, Russell [US/US]; 3761 Bentley Drive, Durham, NC 27707 (US). GUYER, Charles, David [CA/US]; 51 Lake Village Drive, Durham, NC 27713 (US). POTTER, Sharon, Lee [US/US]; 3837 Whispering Branch Road, Raleigh, NC 27613 (US). SUBRAMANIAN, Venkiteswaran [US/US]; 128 Briar Place, Danville, CA 94526 (US). WALTERS, Eric [US/US]; 140 La Cuesta Drive, Scotts Valley, CA 95066 (US).		(74) Agent: BECKER, Konrad; Novartis AG, Corporate Intellectual Property, Patent & Trademark Dept., CH-4002 Basel (CH). (81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). Published <i>Without international search report and to be republished upon receipt of that report.</i> <i>With an indication in relation to deposited biological material furnished under Rule 13bis separately from the description.</i>
(54) Title: METHODS TO SCREEN HERBICIDAL COMPOUNDS UTILIZING AIR SYNTHETASE FROM ARABIDOPSIS THALIANA (57) Abstract <p>The present invention discloses methods to screen chemicals for herbicidal activity using recombinantly produced enzymes having AIR synthetase activity, and the use thereby to identify herbicidal chemicals to suppress the growth of undesired vegetation. Furthermore, the present invention provides methods for the development of herbicide tolerance in plants, plant tissues, plant seeds, and plant cells using genes encoding enzymes having AIR synthetase activity, and methods of using such transgenic plants to selectively suppress weed growth in crop fields.</p>		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

METHOD TO SCREEN HERBICIDAL COMPOUNDS UTILIZING AIR SYNTHETASE FROM ARABIDOPSIS THALIANA

The invention relates to methods for screening herbicidal compounds which inhibit the enzymatic activity of 5'-phosphoribosyl-5-aminoimidazole (AIR) synthetase, an enzyme involved in *de novo* purine biosynthesis. The invention also relates to the use of thereby identified herbicidal chemicals to control the growth of undesired vegetation. The invention may also be applied to the development of herbicide tolerance in plants, plant tissues, plant seeds, and plant cells.

The AIR synthetase is an enzymatic step in the *de novo* purine biosynthesis pathway, which leads to the synthesis of the purine nucleotides IMP, AMP and GMP. *De novo* purine biosynthesis plays a central role in the nitrogen assimilation pathway and is conserved among bacteria, yeast, *Drosophila* and mammals (Schnorr et al. (1994) *The Plant Journal*, 6: 113-121). The AIR synthetase enzymatic activity corresponds to the fifth step in the pathway and catalyzes the conversion of 5'-phosphoribosyl-N-formylglycinamide (FGAM) to 5'-phosphoribosyl-5-aminoimidazole (AIR). In *E. coli*, this step is carried out by a protein encoded by the *purM* gene. Recently, an *Arabidopsis* c-DNA encoding an enzyme having AIR synthetase activity has been cloned and its sequence has been determined (Senecoff and Meagher (1993) *Plant Physiol.* 102: 387-399; Schnorr et al. (1994) *The Plant Journal*, 6: 113-121).

The use of herbicides to control undesirable vegetation such as weeds in crop fields has become almost a universal practice. The herbicide market exceeds 15 billion dollars annually. Despite this extensive use, weed control remains a significant and costly problem for farmers.

Effective use of herbicides requires sound management. For instance, the time and method of application and stage of weed plant development are critical to getting good weed control with herbicides. Since various weed species are resistant to herbicides, the production of effective new herbicides becomes increasingly important. Novel herbicides can now be discovered using high-throughput screens that implement recombinant DNA technology. Metabolic enzymes found to be essential to plant growth and development can be recombinantly produced through standard molecular biological techniques and utilized as herbicide targets in screens for novel inhibitors of the enzymes' activity. The novel

- 2 -

inhibitors discovered through such screens may then be used as herbicides to control undesirable vegetation.

Herbicides that exhibit greater potency, broader weed spectrum, and more rapid degradation in soil can also, unfortunately, have greater crop phytotoxicity. One solution applied to this problem has been to develop crops that are resistant or tolerant to herbicides. Crop hybrids or varieties tolerant to the herbicides allow for the use of the herbicides to kill weeds without attendant risk of damage to the crop. Development of tolerance can allow application of a herbicide to a crop where its use was previously precluded or limited (e.g. to pre-emergence use) due to sensitivity of the crop to the herbicide. For example, U.S. Patent No. 4,761,373 to Anderson *et al.* is directed to plants resistant to various imidazolinone or sulfonamide herbicides. The resistance is conferred by an altered acetohydroxyacid synthase (AHAS) enzyme. U.S. Patent No. 4,975,374 to Goodman *et al.* relates to plant cells and plants containing a gene encoding a mutant glutamine synthetase (GS) resistant to inhibition by herbicides that were known to inhibit GS, e.g. phosphinothricin and methionine sulfoximine. U.S. Patent No. 5,013,659 to Bedbrook *et al.* is directed to plants expressing a mutant acetolactate synthase that renders the plants resistant to inhibition by sulfonylurea herbicides. U.S. Patent No. 5,162,602 to Somers *et al.* discloses plants tolerant to inhibition by cyclohexanedione and aryloxyphenoxypropanoic acid herbicides. The tolerance is conferred by an altered acetyl coenzyme A carboxylase (ACCase).

One object of the present invention is to provide methods for identifying new or improved herbicides. Another object of the invention is to provide methods for using such new or improved herbicides to suppress the growth of plants such as weeds. Still another object of the invention is to provide improved crop plants that are tolerant to such new or improved herbicides.

Using an antisense validation system which allows for the inactivation of expression of an endogenous gene, the inventors of the present invention have demonstrated that the 5'-phosphoribosyl-5-aminoimidazole (AIR) synthetase activity is essential in plants. This implies that chemicals which inhibit AIR synthetase in plants are likely to have detrimental effects on plants and are potentially good herbicide candidates. The present invention therefore provides methods of using a purified AIR synthetase to identify inhibitors thereof, which can then be used as herbicides to suppress the growth of undesirable vegetation,

e.g. in fields where crops are grown, particularly agronomically important crops such as maize and other cereal crops such as wheat, oats, rye, sorghum, rice, barley, millet, turf and forage grasses, and the like, as well as cotton, sugar cane, sugar beet, oilseed rape, and soybeans.

The present invention discloses for the first time the correct nucleotide sequence of the Arabidopsis AIR synthetase gene. The nucleotide sequence encoding the pre-protein is set forth in SEQ ID NO:1 and the nucleotide sequence encoding the putative mature protein is set forth in SEQ ID NO:3. The correct amino acid sequence of the Arabidopsis AIR synthetase pre-protein is set forth in SEQ ID NO:2 and of the correct amino acid sequence of the putative mature Arabidopsis AIR synthetase is set forth in SEQ ID NO:4. The present invention also encompasses isolated enzymes having AIR synthetase activity and comprising an amino acid sequence that is identical or substantially similar to the amino acid sequences set forth in SEQ ID NO:2 or SEQ ID NO:4. Preferably, the amino acid sequence is derived from a plant.

The present invention also encompasses an isolated nucleic acid molecule comprising a nucleotide sequence that encodes the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4. Preferably, the nucleotide sequence is SEQ ID NO:1 or SEQ ID NO:3. In another embodiment, the nucleotide sequence is deposited in *E. coli* strain DH5apASM designated as NRRL accession number B-21976. Also encompassed by the present invention are a chimeric gene comprising a heterologous promoter sequence operatively linked to the nucleic acid molecule of the invention; a recombinant vector comprising such a chimeric gene; and a host cell comprising such a chimeric gene. Preferably, the host cell is a bacterial cell, a yeast cell, or a plant cell. The present invention also encompasses a plant comprising a plant cell of the invention and seed from such a plant.

In a preferred embodiment, the present invention describes a method of identifying chemicals having the ability to inhibit plant growth or viability, comprising: (a) combining an enzyme having AIR synthetase activity in a first reaction mixture with a substrate of AIR synthetase under conditions in which the enzyme is capable of catalyzing the synthesis of AIR; (b) combining the chemical to be tested and the enzyme in a second reaction mixture with a substrate of AIR synthetase under the same conditions and for the same period of time as in the first reaction mixture; (c) determining and comparing the activity of the enzyme in the first and second reaction mixtures; wherein less, desirably significantly less, enzyme activity in the second reaction mixture than in the first reaction mixture indicates

that the chemical of (b) has the ability to inhibit plant growth or viability. In a preferred embodiment, the substrate of AIR synthetase is 5'-phosphoribosyl-N-formylglycinamide (FGAM) and in a further preferred embodiment, the substrate of AIR synthetase is b-FGAM. In another preferred embodiment, the enzyme having AIR synthetase activity is derived from a plant and more preferably, is encoded by a nucleotide sequence identical or substantially similar to the nucleotide sequence set forth in SEQ ID NO:1 or SEQ ID NO:3. In another embodiment, the AIR synthetase enzyme is encoded by a nucleotide sequence capable of encoding the amino acid sequence of SEQ ID NO:2 or SEQ ID NO:4. In yet another embodiment, the AIR synthetase enzyme has an amino acid sequence identical or substantially similar to the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4. In another preferred embodiment, the chemical is capable of inhibiting the growth or viability of a plant by inhibiting the activity of AIR synthetase in the plant. In yet another preferred embodiment, the activity of the enzyme is determined by measuring the AIR produced in the reaction mixture. In another preferred embodiment, the activity of the enzyme is determined by measuring the ADP derived from ATP in the reaction mixture.

In another preferred embodiment, the present invention describes a method of identifying chemicals having the ability to inhibit plant growth or viability, comprising: (a) combining an enzyme having 5'-phosphoribosyl-N-formylglycinamide (FGAM) synthetase activity and an enzyme having AIR synthetase activity in a first reaction mixture with a substrate of FGAM synthetase under conditions in which the enzymes are capable of catalyzing the coupled synthesis of AIR; (b) combining a chemical to be tested and the enzymes in a second reaction mixture with a substrate of FGAM synthetase under the same conditions and the same period of time as in the first reaction mixture; and (c) determining and comparing the activity of the enzyme having AIR synthetase activity in the first and second reaction mixtures; wherein less, preferably significantly less, AIR synthetase enzyme activity in the second reaction mixture than in the first reaction mixture indicates that the chemical of (b) has the ability to inhibit plant growth or viability. In a preferred embodiment, the substrate of FGAM synthetase is 5'-phosphoribosyl-N-formylglycinamide (FGAR) and in a further preferred embodiment, the substrate of FGAM synthetase is b-FGAR. In another preferred embodiment, the enzyme having AIR synthetase activity is derived from a plant and more preferably, is encoded by a nucleotide sequence identical or substantially similar to the nucleotide sequence set forth in SEQ ID NO:1 or SEQ ID NO:3. In another embodiment, the AIR synthetase enzyme is encoded by a nucleotide sequence capable of encoding the amino acid sequence of SEQ ID NO:2 or SEQ ID NO:4. In yet

another embodiment, the AIR synthetase enzyme has an amino acid sequence identical or substantially similar to the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4. In another preferred embodiment, the chemical is capable of inhibiting the growth or viability of a plant by inhibiting the activity of AIR synthetase in the plant. In yet another preferred embodiment, the activity of the enzyme is determined by measuring the AIR produced in the reaction mixture. In another preferred embodiment, the activity of the enzyme is determined by measuring the ADP derived from ATP in the reaction mixture.

The present invention also further describes an assay comprising the steps of: (a) combining an enzyme having 5'-phosphoribosyl-N-formylglycinamide (FGAM) synthetase activity and an enzyme having AIR synthetase activity in a first reaction mixture with a substrate of FGAM synthetase under conditions in which the enzymes are capable of catalyzing the coupled synthesis of AIR; (b) combining a chemical and the enzymes in a second reaction mixture with a substrate of FGAM synthetase under the same conditions and for the same period of time as in the first reaction mixture; (c) determining the activity of the enzyme having AIR synthetase activity in the first and second reaction mixtures; wherein the chemical is capable of inhibiting the activity of the enzyme having AIR synthetase activity if the activity of the enzyme having AIR synthetase activity in the second reaction mixture is less, desirably significantly less, than the activity of the enzyme having AIR synthetase activity in the first reaction mixture. In a preferred embodiment, the substrate of FGAM synthetase is 5'-phosphoribosyl-N-formylglycinamide (FGAR) and in a further preferred embodiment, the substrate of FGAM synthetase is b-FGAR. In yet another preferred embodiment, the activity of the enzyme is determined by measuring the AIR produced in the reaction mixture. In another preferred embodiment, the reaction mixture comprises ATP and the activity of the enzyme is determined by measuring the ADP derived from ATP in the reaction mixture.

In another preferred embodiment, the present invention describes a method for identifying chemicals having herbicidal activity that inhibit AIR synthetase activity in plants, comprising: (a) obtaining transgenic plants, plant tissue, plant seeds or plant cells comprising an isolated nucleotide sequence encoding an enzyme having AIR synthetase activity and capable of overexpressing an enzymatically active AIR synthetase; (b) applying a chemical to be tested to the transgenic plants, plant cells, tissues or parts and to the isogenic non-transformed plants, plant cells, tissues or parts; (c) determining the growth or viability of the transgenic and non-transformed plants, plant cells, tissues after application of the chemical; and (d) comparing the growth or viability of the transgenic and non-

- 6 -

transformed plants, plant cells, tissues after application of the chemical; wherein suppression of the growth or viability of the non-transgenic plants, plant cells, tissues or parts, without significantly suppressing the growth or viability of the isogenic transgenic plants, plant cells, tissues or parts indicates that the chemical of (b) has herbicidal activity that inhibits AIR synthetase activity in plants. Desirably, the chemical suppresses the viability or growth of the non-transgenic plants, plant cells, tissues or parts, without significantly suppressing the growth or viability of the isogenic transgenic plants, plant cells, tissues or parts. In a preferred embodiment, the enzyme having AIR synthetase activity is encoded by a nucleotide sequence identical or substantially similar to the nucleotide sequence set forth in SEQ ID NO:1 or SEQ ID NO:3. In another embodiment, the AIR synthetase enzyme is encoded by a nucleotide sequence capable of encoding the amino acid sequence of SEQ ID NO:2 or SEQ ID NO:4. In yet another embodiment, the AIR synthetase enzyme has an amino acid sequence identical or substantially similar to the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4.

The present invention further embodies plants, plant tissues, plant seeds, and plant cells that have modified AIR synthetase activity and that are therefore tolerant to inhibition by a herbicide at levels normally inhibitory to naturally occurring AIR synthetase activity. Herbicide tolerant plants encompassed by the invention include those that would otherwise be potential targets for normally inhibiting herbicides, particularly the agronomically important crops mentioned above. According to this embodiment, plants, plant tissue, plant seeds, or plant cells are transformed, preferably stably transformed, with a recombinant DNA molecule comprising a suitable promoter functional in plants operatively linked to a nucleotide coding sequence that encodes a modified AIR synthetase that is tolerant to inhibition by a herbicide at a concentration that would normally inhibit the activity of wild-type, unmodified AIR synthetase. Modified AIR synthetase activity may also be conferred upon a plant by increasing expression of wild-type herbicide-sensitive AIR synthetase by providing multiple copies of wild-type AIR synthetase genes to the plant or by overexpression of wild-type AIR synthetase genes under control of a stronger-than-wild-type promoter. The transgenic plants, plant tissue, plant seeds, or plant cells thus created are then selected by conventional selection techniques, whereby herbicide tolerant lines are isolated, characterized, and developed. Alternately, random or site-specific mutagenesis may be used to generate herbicide tolerant lines.

Therefore, the present invention provides a plant, plant cell, plant seed, or plant tissue transformed with a DNA molecule comprising a nucleotide sequence isolated from a

plant that encodes an enzyme having AIR synthetase activity, wherein the enzyme has AIR synthetase activity and wherein the DNA molecule confers upon the plant, plant cell, plant seed, or plant tissue tolerance to a herbicide in amounts that normally inhibits naturally occurring AIR synthetase activity. According to one example of this embodiment, the enzyme having AIR synthetase activity is encoded by a nucleotide sequence identical or substantially similar to the nucleotide sequence set forth in SEQ ID NO:1 or SEQ ID NO:3, or has an amino acid sequence identical or substantially similar to the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4.

The invention also provides a method for suppressing the growth of a plant comprising the step of applying to the plant a chemical that inhibits the naturally occurring AIR synthetase activity in the plant. In a related aspect, the present invention is directed to a method for selectively suppressing the growth of weeds in a field containing a crop of planted crop seeds or plants, comprising the steps of: (a) planting herbicide tolerant crops or crop seeds, which are plants or plant seeds that are tolerant to a herbicide that inhibits the naturally occurring AIR synthetase activity; and (b) applying to the crops or crop seeds and the weeds in the field a herbicide in amounts that inhibit naturally occurring AIR synthetase activity, wherein the herbicide suppresses the growth of the weeds without significantly suppressing the growth of the crops.

The present invention further provides a method for forming a mutagenized DNA molecule encoding an enzyme having AIR synthetase activity from a template DNA molecule encoding an enzyme having AIR synthetase activity, wherein said template DNA molecule has been cleaved into double-stranded-random fragments, comprising the steps of: (a) adding to the resultant population of double-stranded-random fragments at least one single-stranded or double-stranded oligonucleotide, wherein said oligonucleotide comprises an area of identity and an area of heterology to the template DNA molecule; (b) denaturing the resultant mixture of double-stranded-random fragments and oligonucleotides into single-stranded molecules; (c) incubating the resultant population of single-stranded molecules with a polymerase under conditions which result in the annealing of said single-stranded molecules at said areas of identity to form pairs of annealed fragments, said areas of identity being sufficient for one member of a pair to prime replication of the other, thereby forming a mutagenized double-stranded polynucleotide; (d) repeating the second and third steps for at least two further cycles, wherein the resultant mixture in the second step of a further cycle includes the mutagenized double-stranded polynucleotide from the third step of the previous cycle, and the further cycle forms a further mutagenized double-stranded

polynucleotide; wherein the mutagenized double-stranded polynucleotide encodes an AIR synthetase enzyme having enhanced tolerance to a herbicide which inhibits the AIR synthetase activity encoded by the template DNA molecule. Also provided is a mutagenized DNA molecule encoding an enzyme having AIR synthetase activity obtained by the above method, wherein said mutagenized DNA molecule encodes an AIR synthetase enzyme having enhanced tolerance to a herbicide which inhibits the AIR synthetase activity encoded by said template DNA molecule.

The present invention also provides a method for forming a mutagenized DNA molecule encoding an enzyme having AIR synthetase activity from at least two non-identical template DNA molecules encoding enzymes having AIR synthetase activity, comprising the steps of: (a) adding to the template DNA molecules at least one oligonucleotide comprising an area of identity to each of the template DNA molecule; (b) denaturing the resultant mixture into single-stranded molecules; (c) incubating the resultant population of single-stranded molecules with a polymerase under conditions which result in the annealing of the oligonucleotides to the template DNA molecules, wherein the conditions for polymerization by the polymerase are such that polymerization products corresponding to a portion of the template DNA molecules are obtained; (d) repeating the second and third steps for at least two further cycles, wherein the extension products obtained in the third step are able to switch template DNA molecule for polymerization in the next cycle, thereby forming a mutagenized double-stranded polynucleotide comprising sequences derived from different template DNA molecules; wherein the mutagenized double-stranded polynucleotide encodes an AIR synthetase enzyme having enhanced tolerance to a herbicide which inhibits the AIR synthetase activity encoded by the template DNA molecules. Also provided is a mutagenized DNA molecule encoding an enzyme having AIR synthetase activity obtained by the above method, wherein said mutagenized DNA molecule encodes an AIR synthetase enzyme having enhanced tolerance to a herbicide which inhibits the AIR synthetase activity encoded by said template DNA molecule.

Preferably, according to either of the above two methods, at least one template DNA molecule is derived from a eukaryote. More preferably, said eukaryote is a plant. Still more preferably, said plant is *Arabidopsis thaliana*. Most preferably, said species of template DNA molecule is identical or substantially similar to the SEQ ID NO:1 or SEQ ID NO:3. In another embodiment of either of the above two methods, at least one template DNA molecule is derived from a prokaryote.

Other objects and advantages of the present invention will become apparent to those skilled in the art from a study of the following description of the invention and non-limiting examples.

For clarity, certain terms used in the specification are defined and presented as follows:

Activatable DNA Sequence: a DNA sequence that regulates the expression of genes in a genome, desirably the genome of a plant. The activatable DNA sequence is complementary to a target gene endogenous in the genome. When the activatable DNA sequence is introduced and expressed in a cell, it inhibits expression of the target gene. An activatable DNA sequence useful in conjunction with the present invention includes those encoding or acting as dominant inhibitors, such as a translatable or untranslatable sense sequence capable of disrupting gene function in stably transformed plants to positively identify one or more genes essential for normal growth and development of a plant. A preferred activatable DNA sequence is an antisense DNA sequence. The target gene preferably encodes a protein, such as a biosynthetic enzyme, receptor, signal transduction protein, structural gene product, or transport protein that is essential to the growth or survival of the plant. In an especially preferred embodiment, the target gene encodes an enzyme having AIR synthetase activity. The interaction of the antisense sequence and the target gene results in substantial inhibition of the expression of the target gene so as to kill the plant, or at least inhibit normal plant growth or development.

Activatable DNA Construct: a recombinant DNA construct comprising a synthetic promoter operatively linked to the activatable DNA sequence, which when introduced into a cell, desirably a plant cell, is not expressed, i.e. is silent, unless a complete hybrid transcription factor capable of binding to and activating the synthetic promoter is present. The activatable DNA construct is introduced into cells, tissues, or plants to form stable transgenic lines capable of expressing the activatable DNA sequence.

Co-factor: natural reactant, such as an organic molecule or a metal ion, required in an enzyme-catalyzed reaction. A co-factor is e.g. NAD(P), riboflavin (including FAD and FMN), folate, molybdopterin, thiamin, biotin, lipoic acid, pantothenic acid and coenzyme A, S-adenosylmethionine, pyridoxal phosphate, ubiquinone, menaquinone.

Coupled synthesis: a enzymatic biosynthesis, in which a final product is synthesized by two sequential enzymatic steps, wherein the substrate for the first enzymatic step is converted by the first enzyme to an intermediate product, which serves as a substrate for

the second enzymatic step and is converted by the second enzyme to the final product, without external addition of the intermediate product.

DNA shuffling: DNA shuffling is a method to introduce mutations or rearrangements, preferably randomly, in a DNA molecule or to generate exchanges of DNA sequences between two or more DNA molecules, preferably randomly. The DNA molecule resulting from DNA shuffling is a shuffled DNA molecule that is a non-naturally occurring DNA molecule derived from at least one template DNA molecule. The shuffled DNA encodes an enzyme modified with respect to the enzyme encoded by the template DNA, and preferably has an altered biological activity with respect to the enzyme encoded by the template DNA.

Enzyme activity: means herein the ability of an enzyme to catalyze the conversion of a substrate into a product. A substrate for the enzyme comprises the natural substrate of the enzyme but also comprises analogues of the natural substrate which can also be converted by the enzyme into a product or into an analogue of a product. The activity of the enzyme is measured for example by determining the amount of product in the reaction after a certain period of time, or by determining the amount of substrate remaining in the reaction mixture after a certain period of time. The activity of the enzyme is also measured by determining the amount of an unused co-factor of the reaction remaining in the reaction mixture after a certain period of time or by determining the amount of used co-factor in the reaction mixture after a certain period of time. The activity of the enzyme is also measured by determining the amount of a donor of free energy or energy-rich molecule (e.g. ATP, phosphoenolpyruvate, acetyl phosphate or phosphocreatine) remaining in the reaction mixture after a certain period of time or by determining the amount of a used donor of free energy or energy-rich molecule (e.g. ADP, pyruvate, acetate or creatine) in the reaction mixture after a certain period of time.

Herbicide: a chemical substance used to kill or suppress the growth of plants, plant cells, plant seeds, or plant tissues.

Heterologous DNA Sequence: a DNA sequence not naturally associated with a host cell into which it is introduced, including non-naturally occurring multiple copies of a naturally occurring DNA sequence.

Homologous DNA Sequence: a DNA sequence naturally associated with a host cell into which it is introduced.

Inhibitor: a chemical substance that inactivates the enzymatic activity of a protein such as a biosynthetic enzyme, receptor, signal transduction protein, structural gene product, or transport protein that is essential to the growth or survival of the plant. In the

context of the instant invention, an inhibitor is a chemical substance that inactivates the enzymatic activity of AIR synthetase from a plant. The term "herbicide" is used herein to define an inhibitor when applied to plants, plant cells, plant seeds, or plant tissues.

Isogenic: plants which are genetically identical, except that they may differ by the presence or absence of a transgene.

Isolated: in the context of the present invention, an isolated DNA molecule or an isolated enzyme is a DNA molecule or enzyme that, by the hand of man, exists apart from its native environment and is therefore not a product of nature. An isolated DNA molecule or enzyme may exist in a purified form or may exist in a non-native environment such as, for example, a transgenic host cell.

Mature protein: protein which is normally targeted to a cellular organelle, such as a chloroplast, and from which the transit peptide has been removed.

Minimal Promoter: promoter elements, particularly a TATA element, that are inactive or that have greatly reduced promoter activity in the absence of upstream activation. In the presence of a suitable transcription factor, the minimal promoter functions to permit transcription.

Modified Enzyme Activity: enzyme activity different from that which naturally occurs in a plant (i.e. enzyme activity that occurs naturally in the absence of direct or indirect manipulation of such activity by man), which is tolerant to inhibitors that inhibit the naturally occurring enzyme activity.

Pre-protein: protein which is normally targeted to a cellular organelle, such as a chloroplast, and still comprising its transit peptide.

Significant Increase: an increase in enzymatic activity that is larger than the margin of error inherent in the measurement technique, preferably an increase by about 2-fold or greater of the activity of the wild-type enzyme in the presence of the inhibitor, more preferably an increase by about 5-fold or greater, and most preferably an increase by about 10-fold or greater.

Significantly less: means that the amount of a product of an enzymatic reaction is larger than the margin of error inherent in the measurement technique, preferably a decrease by about 2-fold or greater of the activity of the wild-type enzyme in the absence of the inhibitor, more preferably an decrease by about 5-fold or greater, and most preferably an decrease by about 10-fold or greater.

In its broadest sense, the term "substantially similar", when used herein with respect to a nucleotide sequence, means a nucleotide sequence corresponding to a reference

nucleotide sequence, wherein the corresponding sequence encodes a polypeptide having substantially the same structure and function as the polypeptide encoded by the reference nucleotide sequence, e.g. where only changes in amino acids not affecting the polypeptide function occur. Desirably the substantially similar nucleotide sequence encodes the polypeptide encoded by the reference nucleotide sequence. The percentage of identity between the substantially similar nucleotide sequence and the reference nucleotide sequence desirably is at least 65%, more desirably at least 75%, preferably at least 85%, more preferably at least 90%, still more preferably at least 95%, yet still more preferably at least 99%. Sequence comparisons are carried out using a Smith-Waterman sequence alignment algorithm (see e.g. Waterman, M.S. Introduction to Computational Biology: Maps, sequences and genomes. Chapman & Hall. London: 1995. ISBN 0-412-99391-0, or at <http://www.hto.usc.edu/software/seqaln/index.html>). The localS program, version 1.16, is used with following parameters: match: 1, mismatch penalty: 0.33, open-gap penalty: 2, extended-gap penalty: 2. A nucleotide sequence "substantially similar" to reference nucleotide sequence hybridizes to the reference nucleotide sequence in 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄, 1 mM EDTA at 50°C with washing in 2X SSC, 0.1% SDS at 50°C, more desirably in 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄, 1 mM EDTA at 50°C with washing in 1X SSC, 0.1% SDS at 50°C, more desirably still in 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄, 1 mM EDTA at 50°C with washing in 0.5X SSC, 0.1% SDS at 50°C, preferably in 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄, 1 mM EDTA at 50°C with washing in 0.1X SSC, 0.1% SDS at 50°C, more preferably in 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄, 1 mM EDTA at 50°C with washing in 0.1X SSC, 0.1% SDS at 65°C.

The term "substantially similar", when used herein with respect to a protein, means a protein corresponding to a reference protein, wherein the protein has substantially the same structure and function as the reference protein, e.g. where only changes in amino acids sequence not affecting the polypeptide function occur. When used for a protein or an amino acid sequence the percentage of identity between the substantially similar and the reference protein or amino acid sequence desirably is at least 65%, more desirably at least 75%, preferably at least 85%, more preferably at least 90%, still more preferably at least 95%, yet still more preferably at least 99%.

Substrate: a substrate is the molecule that the enzyme naturally recognizes and converts to a product in the biochemical pathway in which the enzyme naturally carries out its function, or is a modified version of the molecule, which is also recognized by the

enzyme and is converted by the enzyme to a product in an enzymatic reaction similar to the naturally-occurring reaction.

Tolerance: the ability to continue normal growth or function when exposed to an inhibitor or herbicide.

Transformation: a process for introducing heterologous DNA into a cell, tissue, or plant. Transformed cells, tissues, or plants are understood to encompass not only the end product of a transformation process, but also transgenic progeny thereof.

Transgenic: stably transformed with a recombinant DNA molecule that preferably comprises a suitable promoter operatively linked to a DNA sequence of interest.

BRIEF DESCRIPTION OF THE SEQUENCES IN THE SEQUENCE LISTING

SEQ ID NO:1	DNA sequence encoding the Arabidopsis AIR synthetase pre-protein
SEQ ID NO:2	amino acid sequence of the Arabidopsis AIR synthetase pre-protein
SEQ ID NO:3	DNA sequence encoding the putative mature Arabidopsis AIR synthetase
SEQ ID NO:4	amino acid sequence of the putative mature Arabidopsis AIR synthetase
SEQ ID NO:5	oligonucleotide JG-L
SEQ ID NO:6	oligonucleotide AS-1
SEQ ID NO:7	oligonucleotide AS-2
SEQ ID NO:8	oligonucleotide slp242
SEQ ID NO:9	oligonucleotide slp244
SEQ ID NO:10	oligonucleotide slp243

DEPOSIT

The following material has been deposited with the Agricultural Research Service, Patent Culture Collection (NRRL), 1815 North University Street, Peoria, Illinois 61604, under the terms of the Budapest Treaty on the International Recognition of the Deposit of Microorganisms for the Purposes of Patent Procedure. All restrictions on the availability of the deposited material will be irrevocably removed upon the granting of a patent.

<u>Clone</u>	<u>Accession number</u>	<u>Date of Deposit</u>
DH5apASM	NRRL B-21976	April 17, 1998

I. Correct Sequence of the Arabidopsis AIR Synthetase Gene

The Arabidopsis AIR synthetase gene was re-sequenced by the inventors of the present invention and compared to a published DNA sequence for the Arabidopsis AIR

synthetase gene (Genbank accession L12457, Senecoff and Meagher (1993) Plant Physiol. 102: 387-399). Sequencing results revealed a substantial error in the published DNA sequence, resulting in the insertion of a cytosine base at the position corresponding to position 1,027 in SEQ ID NO:1. This insertion leads to a frame-shift mutation in the amino acid sequence and therefore teaches away from the correct deduced amino acid sequence for the Arabidopsis AIR synthetase. The present invention discloses for the first time the correct nucleotide sequence of the Arabidopsis AIR synthetase gene as well as the correct amino acid sequence of the Arabidopsis AIR synthetase. The nucleotide sequence encoding the pre-protein is set forth in SEQ ID NO:1 and the nucleotide sequence encoding the mature protein is set forth in SEQ ID NO:3. The correct amino acid sequence of the Arabidopsis AIR synthetase pre-protein encoded by the nucleotide sequence set forth in SEQ ID NO:1 is set forth in SEQ ID NO:2 and the correct amino acid sequence of the putative mature Arabidopsis AIR synthetase encoded by the nucleotide sequence set forth in SEQ ID NO:2 is set forth in SEQ ID NO:4. The nucleotide sequence encoding the Arabidopsis AIR synthetase pre-protein was deposited in *E. coli* strain DH5apASM and designated as NRRL accession number B-21976. The present invention also encompasses an isolated amino acid sequence derived from a plant, wherein said amino acid sequence is identical or substantially similar to the amino acid sequence encoded by the nucleotide sequence set forth in SEQ ID NO:1 or SEQ ID NO:3, wherein said amino acid sequence has 5'-phosphoribosyl-5-aminoimidazole (AIR) synthetase activity. The present invention also further encompasses an isolated amino acid sequence derived from a plant, wherein said amino acid sequence is identical or substantially similar to the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4, wherein said amino acid sequence has 5'-phosphoribosyl-5-aminoimidazole (AIR) synthetase activity.

II. Essentiality of the AIR Synthetase Gene in Plants Demonstrated by Antisense Inhibition

As shown in the examples below, the essentiality of the AIR synthetase gene for normal plant growth and development has been demonstrated for the first time by antisense inhibition in plants using the antisense validation system described in PCT application no. EP98/07577, incorporated herein by reference. Having established the essentiality of AIR synthetase function in plants, the inventors thereby provide an important and sought after tool for new herbicide development.

In the system described in the present invention, a hybrid transcription factor gene is made that comprises a DNA-binding domain and an activation domain. In addition, an activatable DNA construct is made that comprises a synthetic promoter operatively linked to an activatable DNA sequence. The hybrid transcription factor gene and synthetic promoter are selected or designed such that the DNA binding domain of the hybrid transcription factor is capable of binding specifically to the synthetic promoter, which then activates expression of the activatable DNA sequence. A first plant is transformed with the hybrid transcription factor gene, and a second plant is transformed with the activatable DNA construct. The first plant and second plants are crossed to produce a progeny plant containing both the sequence encoding the hybrid transcription factor and the synthetic promoter, wherein the activatable DNA sequence is expressed in the progeny plant. In the preferred embodiment, the activatable DNA sequence is an antisense sequence capable of inactivating expression of an endogenous gene such as the AIR synthetase gene. Hence, the progeny plant will be unable to normally express the endogenous gene.

This antisense validation system is especially useful for allowing expression of traits that might otherwise be unrecoverable as constitutively driven transgenes. For instance, foreign genes with potentially lethal effect or antisense genes or dominant-negative mutations designed to abolish function of essential genes, while of great interest in basic studies of plant biology, present inherent experimental problems. Decreased transformation frequencies are often cited as evidence of lethality associated with a particular constitutively driven transgene, but negative results of this type are laden with alternative trivial explanations. The present invention is an important advancement in the field of agriculture because it allows stable maintenance and propagation of a test transgene separate from its expression. This ability to separate transgene insertion from expression is especially useful for firm conclusions about essentiality of gene function to be drawn. A substantial benefit of the present invention is that plant genes essential for normal growth or development can thus be identified in this manner. The identification of such genes provide useful targets for screening compound libraries for identification of effective herbicides. Below, the antisense validation system is described in greater detail:

A. Hybrid Transcription Factor Gene

A hybrid transcription factor gene for use in the antisense validation system described herein comprises DNA sequences encoding (1) a DNA-binding domain and (2) an activation domain that interacts with components of transcriptional machinery assembling at

a promoter. Gene fragments are joined, typically such that the DNA binding domain is toward the 5' terminus and the activator domain is toward the 3' terminus, to form a hybrid gene whose expression produces a hybrid transcription factor. One skilled in the art is capable of routinely combining various DNA sequences encoding DNA binding domains with various DNA sequences encoding activation domains to produce a wide array of hybrid transcription factor genes. Examples of DNA sequences encoding DNA-binding domains include, but are not limited to, those encoding the DNA binding domains of GAL4, bacteriophage 434, *lexA*, *lacI*, and phage lambda repressor. Examples of DNA sequences encoding the activation domain include, but are not limited to, those encoding the acidic activation domains of herpes simplex VP16, maize C1, and P1. In addition, suitable activation domains can be isolated by fusing DNA pieces from an organism of choice to a suitable DNA binding domain and selecting directly for function (Estruch *et al.*, (1994) *Nucleic Acids Res.* 22: 3983-3989). Domains of transcriptional activator proteins can be swapped between proteins of diverse origin (Brent and Ptashne (1985) *Cell* 43: 729-736). A preferable hybrid transcription factor gene comprises DNA sequences encoding the GAL4 DNA binding domain fused to the maize C1 activation domain.

B. Activatable DNA Construct

An activatable DNA construct for use in the antisense validation system described herein comprises (1) a synthetic promoter operatively linked to (2) an activatable DNA sequence. The synthetic promoter comprises at least one DNA binding site recognized by the DNA binding domain of the hybrid transcription factor, and a minimal promoter, preferably a TATA element derived from a promoter recognized by plant cells. More particularly the TATA element is derived from a promoter recognized by the plant cell type into which the synthetic promoter will be incorporated. Desirably, the DNA binding site is repeated multiple times in the synthetic promoter so that the minimal promoter may be more effectively activated, such that the activatable DNA sequence associated with the synthetic promoter is more effectively expressed. One skilled in the art can use routine molecular biology and recombinant DNA technology to make desirable synthetic promoters. Examples of DNA binding sites that can be used to make synthetic promoters useful in the invention include, but are not limited to, the upstream activating sequence (UAS_G) recognized by the GAL4 DNA binding protein, the *lac* operator, and the *lexA* binding site. Examples of promoter TATA elements recognized by plant cells include those derived from CaMV 35S, the maize *Bz1* promoter, and the UBQ3 promoter. An especially preferable

- 17 -

synthetic promoter comprises a truncated CaMV 35S sequence containing the TATA element (nucleotides -59 to +48 relative to the start of transcription), fused at its 5' end to approximately 10 concatemeric direct repeats of the upstream activating sequence (UAS_G) recognized by the GAL4 DNA binding domain.

The activatable DNA sequence encompasses any DNA sequence for which stable introduction and expression in a plant cell is desired. Particularly desirable activatable DNA sequences are sense or antisense sequences, whose expression results in decreased expression of their endogenous counterpart genes, thereby inhibiting normal plant growth or development. The activatable DNA sequence is operatively linked to the synthetic promoter to form the activatable DNA construct. The activatable DNA sequence in the activatable DNA construct is not expressed, i.e. is silent, in transgenic lines, unless a hybrid transcription factor capable of binding to and activating the synthetic promoter, is also present. The activatable DNA construct subsequently is introduced into cells, tissues or plants to form stable transgenic lines expressing the activatable DNA sequence, as described more fully below. In the context of the present invention, the activatable DNA sequence preferably comprises an antisense AIR synthetase sequence.

C. Transgenic Plants Containing the Hybrid Transcription Factor Gene or the Activatable DNA Construct

The antisense validation system described herein utilizes a first plant containing the hybrid transcription factor gene and a second plant containing the activatable DNA construct. The hybrid transcription factor genes and activatable DNA constructs described above are introduced into the plants by methods well known and routinely used in the art, including but not limited to crossing, *Agrobacterium*-mediated transformation, Ti plasmid vectors, direct DNA uptake such as microprojectile bombardment, liposome mediated uptake, micro-injection, etc. Transformants are screened for the presence and functionality of the transgenes according to standard methods known to those skilled in the art.

D. Transgenic Plants Containing Both the Hybrid Transcription Factor Gene and the Activatable DNA Construct

F1 plants containing both the hybrid transcription factor gene and the activatable DNA construct are generated by cross-pollination and selected for the presence of an appropriate marker. In contrast to plants containing the activatable DNA construct alone,

the F1 plants generate high levels of activatable DNA sequence expression product, comparable to those obtained with strong constitutive promoters such as CaMV 35S.

E. Antisense Validation Assay

Thus, a useful assay in the system described herein comprises the following steps:

providing a first transgenic plant stably transformed with a hybrid transcription factor gene encoding a hybrid transcription factor capable of activating a synthetic promoter when said synthetic promoter is present in the plant, wherein the first transgenic plant is homozygous for the hybrid transcription factor; b) providing a second transgenic plant stably transformed with an activatable DNA construct comprising a synthetic promoter activatable by the hybrid transcription factor of step a) operatively linked to an activatable DNA sequence, such as an antisense AIR synthetase sequence; c) crossing the first transgenic plant with the second transgenic plant to yield F1 plants expressing the activatable DNA sequence in the presence of the hybrid transcription factor; and d) determining the effect of expression of the activatable DNA sequence on the F1 plants.

III. Recombinant Production of AIR Synthetases and Uses Thereof

For recombinant production of AIR synthetase in a host organism, a nucleotide sequence encoding an enzyme having AIR synthetase is inserted into an expression cassette designed for the chosen host and introduced into the host where it is recombinantly produced. The choice of specific regulatory sequences such as promoter, signal sequence, 5' and 3' untranslated sequences, and enhancer appropriate for the chosen host is within the level of skill of the routineer in the art. The resultant molecule, containing the individual elements linked in proper reading frame, may be inserted into a vector capable of being transformed into the host cell. Suitable expression vectors and methods for recombinant production of proteins are well known for host organisms such as *E. coli*, yeast, and insect cells (see, e.g., Luckow and Summers, *Bio/Technol.* 6: 47 (1988)). Specific examples include plasmids such as pBluescript (Stratagene, La Jolla, CA), pFLAG (International Biotechnologies, Inc., New Haven, CT), pTrcHis (Invitrogen, La Jolla, CA), and baculovirus expression vectors, e.g., those derived from the genome of *Autographica californica* nuclear polyhedrosis virus (AcMNPV). A preferred baculovirus/insect system is pVI11392/Sf21 cells (Invitrogen, La Jolla, CA).

In a preferred embodiment, the nucleotide sequence encoding an enzyme having a AIR synthetase activity is derived from an eukaryote, such as a mammal, a fly or a yeast,

but is preferably derived from a plant. In a further preferred embodiment, the nucleotide sequence is identical or substantially similar to the nucleotide sequence set forth in SEQ ID NO:1 or SEQ ID NO:3, or encodes an enzyme having AIR synthetase activity, whose amino acid sequence is identical or substantially similar to the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4. The nucleotide sequence set forth in SEQ ID NO:1 encodes the Arabidopsis AIR synthetase pre-protein, whose amino acid sequence is set forth in SEQ ID NO:2, and the nucleotide sequence set forth in SEQ ID NO:3 encodes the Arabidopsis putative mature AIR synthetase, whose amino acid sequence is set forth in SEQ ID NO:4. In another preferred embodiment, the nucleotide sequence is derived from a prokaryote, preferably a bacteria, e.g. *E. coli*. In this case, the enzyme having AIR synthetase activity is encoded by the *purM* gene.

Recombinantly produced AIR synthetases are isolated and purified using a variety of standard techniques. The actual techniques that may be used will vary depending upon the host organism used, whether the enzyme is designed for secretion, and other such factors familiar to the skilled artisan (see, e.g. chapter 16 of Ausubel, F. *et al.*, "Current Protocols in Molecular Biology", pub. by John Wiley & Sons, Inc. (1994).

Recombinantly produced AIR synthetases are useful for a variety of purposes. For example, they can be used in *in vitro* assays to screen known herbicidal chemicals whose target has not been identified to determine if they inhibit AIR synthetases. Such *in vitro* assays may also be used as more general screens to identify chemicals that inhibit such enzymatic activity and that are therefore novel herbicide candidates. Alternatively, recombinantly produced AIR synthetases may be used to elucidate the complex structure of these molecules and to further characterize their association with known inhibitors in order to rationally design new inhibitory herbicides as well as herbicide tolerant forms of the enzymes.

IV. *In Vitro* Inhibitor Assay

An *in vitro* assay useful for identifying inhibitors of enzymes encoded by essential plant genes, such as the AIR synthetase, preferably comprises the steps of: a) reacting an enzyme having AIR synthetase activity and a substrate thereof in the presence of a suspected inhibitor of the enzyme's function; b) comparing the rate of enzymatic activity in the presence of the suspected inhibitor to the rate of enzymatic activity under the same conditions in the absence of the suspected inhibitor; and c) determining whether the

suspected inhibitor inhibits the AIR synthetase enzymatic activity. In a preferred embodiment, such a determination is made by comparing, in the presence and absence of the candidate inhibitor, the amount of AIR synthesized in the *in vitro* assay using fluorescence or absorbance detection. In another preferred embodiment, such a determination is made by comparing, in the presence and absence of the candidate inhibitor, the amount of ADP formed in the *in vitro* assay using fluorescence or absorbance detection. A preferred substrate for AIR synthetase is 5'-phosphoribosyl-N-formylglycinamide (FGAM), in particular the b isomer, b-FGAM.

In another preferred embodiment, a coupled FGAM synthetase/AIR synthetase assay is used, thereby increasing the detection limit of the assay and resulting in an improved screening procedure for a chemical inhibiting AIR synthetase activity. Such a coupling assay preferably comprises the steps of: a) reacting an enzyme having 5'-phosphoribosyl-N-formylglycinamide (FGAM) synthetase activity, an enzyme having AIR synthetase activity and a substrate of FGAM synthetase in the presence of a suspected inhibitor of the enzyme's function; b) comparing the rate of enzymatic activity in the presence of the suspected inhibitor to the rate of enzymatic activity under the same conditions in the absence of the suspected inhibitor; and c) determining whether the suspected inhibitor inhibits the AIR synthetase enzymatic activity. In a preferred embodiment, such a determination is made by comparing, in the presence and absence of the candidate inhibitor, the amount of AIR synthesized in the *in vitro* assay using fluorescence or absorbance detection. In another preferred embodiment, such a determination is made by comparing, in the presence and absence of the candidate inhibitor, the amount of ADP formed in the *in vitro* assay using fluorescence or absorbance detection. A preferred substrate for FGAM synthetase is 5'-phosphoribosyl-N-formylglycinamide (FGAR), in particular the b isomer, b-FGAR. In a further preferred embodiment, the enzyme having FGAM synthetase activity is derived from a bacteria, and is preferably the *E. coli* FGAM synthetase encoded by the purL gene. The purL gene is preferably recombinantly produced in *E. coli*. While any suitable AIR synthetase may be used, preferably the AIR synthetase used in such *in vitro* assays is derived from a plant. In another preferred embodiment, an assay coupling more than one enzymatic activity preceding AIR synthetase in the purine biosynthesis pathway is used.

In a preferred embodiment, an enzyme used in an *in vitro* assay is derived from cells comprising the enzyme, preferably, from a crude extract of the cells. The enzyme is

preferably isolated and purified from the cells or from the crude extract. The enzyme is preferably produced recombinantly and is preferably isolated and purified prior to be used in the assay. Chemicals identified in an *in vitro* assay are then tested for their ability to inhibit plant growth or viability.

V. *In Vivo* Inhibitor Assay

A. In one embodiment, a suspected herbicide, for example identified by *in vitro* screening, is applied to plants at various concentrations. The suspected herbicide is preferably sprayed on the plants. After application of the suspected herbicide, its effect on the plants, for example death or suppression of growth is recorded.

B. In another embodiment, an *in vivo* screening assay for inhibitors of the AIR synthetase activity uses transgenic plants, plant tissue, plant seeds or plant cells capable of overexpressing a nucleotide sequence having AIR synthetase activity, wherein the AIR synthetase is enzymatically active in the transgenic plants, plant tissue, plant seeds or plant cells. The nucleotide sequence is preferably derived from an eukaryote, such as a mammal, a fly or a yeast, but is preferably derived from a plant. In a further preferred embodiment, the nucleotide sequence is identical or substantially similar to the nucleotide sequence set forth in SEQ ID NO:1 or SEQ ID NO:3, or encodes an enzyme having AIR synthetase activity, whose amino acid sequence is identical or substantially similar to the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4. In another preferred embodiment, the nucleotide sequence is derived from a prokaryote, preferably a bacteria, e.g. *E. coli*. In this case, the enzyme having AIR synthetase activity is encoded by the *purM* gene.

A chemical is then applied to the transgenic plants, plant tissue, plant seeds or plant cells and to the isogenic non-transformed plants, plant tissue, plant seeds or plant cells, and the growth or viability of the transgenic and non-transformed plants, plant tissue, plant seeds or plant cells are determined after application of the chemical and compared.

VI. Herbicide Tolerant Plants

The present invention is further directed to plants, plant tissue, plant seeds, and plant cells tolerant to herbicides that inhibit the naturally occurring AIR synthetase activity in these plants, wherein the tolerance is conferred by an altered AIR synthetase activity. Altered AIR synthetase activity may be conferred upon a plant according to the invention by increasing expression of wild-type herbicide-sensitive AIR synthetase by providing additional wild-type AIR synthetase genes to the plant, by expressing modified

herbicide-tolerant AIR synthetases in the plant, or by a combination of these techniques. Representative plants include any plants to which these herbicides are applied for their normally intended purpose. Preferred are agronomically important crops such as cotton, soybean, oilseed rape, sugar beet, maize, rice, wheat, barley, oats, rye, sorghum, millet, turf, forage, turf grasses, and the like.

A. Increased Expression of Wild-Type AIR Synthetase

Achieving altered AIR synthetase activity through increased expression results in a level of a AIR synthetase in the plant cell at least sufficient to overcome growth inhibition caused by the herbicide. The level of expressed enzyme generally is at least two times, preferably at least five times, and more preferably at least ten times the natively expressed amount. Increased expression may be due to multiple copies of a wild-type AIR synthetase gene; multiple occurrences of the coding sequence within the gene (*i.e.* gene amplification) or a mutation in the non-coding, regulatory sequence of the endogenous gene in the plant cell. Plants having such altered gene activity can be obtained by direct selection in plants by methods known in the art (see, *e.g.* U.S. Patent No. 5,162,602, and U.S. Patent No. 4,761,373, and references cited therein). These plants also may be obtained by genetic engineering techniques known in the art. Increased expression of a herbicide-sensitive AIR synthetase gene can also be accomplished by transforming a plant cell with a recombinant or chimeric DNA molecule comprising a promoter capable of driving expression of an associated structural gene in a plant cell operatively linked to a homologous or heterologous structural gene encoding the AIR synthetase. Preferably, the transformation is stable, thereby providing a heritable transgenic trait.

B. Expression of Modified Herbicide-Tolerant AIR Synthetases

According to this embodiment, plants, plant tissue, plant seeds, or plant cells are stably transformed with a recombinant DNA molecule comprising a suitable promoter functional in plants operatively linked to a coding sequence encoding a herbicide tolerant form of an AIR synthetase. A herbicide tolerant form of the enzyme has at least one amino acid substitution, addition or deletion that confers tolerance to a herbicide that inhibits the unmodified, naturally occurring form of the enzyme. The transgenic plants, plant tissue, plant seeds, or plant cells thus created are then selected by conventional selection techniques, whereby herbicide tolerant lines are isolated, characterized, and developed.

Below are described methods for obtaining genes that encode herbicide tolerant forms of AIR synthetases:

One general strategy involves direct or indirect mutagenesis procedures on microbes. For instance, a genetically manipulatable microbe such as *E. coli* or *S. cerevisiae* may be subjected to random mutagenesis *in vivo* with mutagens such as UV light or ethyl or methyl methane sulfonate. Mutagenesis procedures are described, for example, in Miller, *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1972); Davis *et al.*, *Advanced Bacterial Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1980); Sherman *et al.*, *Methods in Yeast Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1983); and U.S. Patent No. 4,975,374. The microbe selected for mutagenesis contains a normal, inhibitor-sensitive AIR synthetase gene and is dependent upon the activity conferred by this gene. The mutagenized cells are grown in the presence of the inhibitor at concentrations that inhibit the unmodified gene. Colonies of the mutagenized microbe that grow better than the unmutagenized microbe in the presence of the inhibitor (i.e. exhibit resistance to the inhibitor) are selected for further analysis. AIR synthetase genes from these colonies are isolated, either by cloning or by PCR amplification, and their sequences are elucidated. Sequences encoding altered gene products are then cloned back into the microbe to confirm their ability to confer inhibitor tolerance.

A method of obtaining mutant herbicide-tolerant alleles of a plant AIR synthetase gene involves direct selection in plants. For example, the effect of a mutagenized AIR synthetase gene on the growth inhibition of plants such as *Arabidopsis*, soybean, or maize is determined by plating seeds sterilized by art-recognized methods on plates on a simple minimal salts medium containing increasing concentrations of the inhibitor. Such concentrations are in the range of 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 110, 300, 1000 and 3000 parts per million (ppm). The lowest dose at which significant growth inhibition can be reproducibly detected is used for subsequent experiments. Determination of the lowest dose is routine in the art.

Mutagenesis of plant material is utilized to increase the frequency at which resistant alleles occur in the selected population. Mutagenized seed material is derived from a variety of sources, including chemical or physical mutagenesis of seeds, or chemical or physical mutagenesis of pollen (Neuffer, In *Maize for Biological Research* Sheridan, ed. Univ. Press, Grand Forks, ND., pp. 61-64 (1982)), which is then used to fertilize plants and

the resulting M₁ mutant seeds collected. Typically for *Arabidopsis*, M₂ seeds (Lehle Seeds, Tucson, AZ), which are progeny seeds of plants grown from seeds mutagenized with chemicals, such as ethyl methane sulfonate, or with physical agents, such as gamma rays or fast neutrons, are plated at densities of up to 10,000 seeds/plate (10 cm diameter) on minimal salts medium containing an appropriate concentration of inhibitor to select for tolerance. Seedlings that continue to grow and remain green 7-21 days after plating are transplanted to soil and grown to maturity and seed set. Progeny of these seeds are tested for tolerance to the herbicide. If the tolerance trait is dominant, plants whose seed segregate 3:1 / resistant:sensitive are presumed to have been heterozygous for the resistance at the M₂ generation. Plants that give rise to all resistant seed are presumed to have been homozygous for the resistance at the M₂ generation. Such mutagenesis on intact seeds and screening of their M₂ progeny seed can also be carried out on other species, for instance soybean (see, e.g. U.S. Pat. No. 5,084,082). Alternatively, mutant seeds to be screened for herbicide tolerance are obtained as a result of fertilization with pollen mutagenized by chemical or physical means.

Confirmation that the genetic basis of the herbicide tolerance is a modified AIR synthetase gene is ascertained as exemplified below. First, alleles of the AIR synthetase gene from plants exhibiting resistance to the inhibitor are isolated using PCR with primers based either upon the *Arabidopsis* cDNA coding sequences shown in SEQ ID NO:1 or, more preferably, based upon the unaltered AIR synthetase gene sequence from the plant used to generate tolerant alleles. After sequencing the alleles to determine the presence of mutations in the coding sequence, the alleles are tested for their ability to confer tolerance to the inhibitor on plants into which the putative tolerance-conferring alleles have been transformed. These plants can be either *Arabidopsis* plants or any other plant whose growth is susceptible to the AIR synthetase inhibitors. Second, the inserted AIR synthetase genes are mapped relative to known restriction fragment length polymorphisms (RFLPs) (See, for example, Chang *et al. Proc. Natl. Acad. Sci. USA* 85: 6856-6860 (1988); Nam *et al., Plant Cell* 1: 699-705 (1989). The AIR synthetase inhibitor tolerance trait is independently mapped using the same markers. When tolerance is due to a mutation in that AIR synthetase gene, the tolerance trait maps to a position indistinguishable from the position of the AIR synthetase gene.

Another method of obtaining herbicide-tolerant alleles of a AIR synthetase gene is by selection in plant cell cultures. Explants of plant tissue, e.g. embryos, leaf disks, etc. or

actively growing callus or suspension cultures of a plant of interest are grown on medium in the presence of increasing concentrations of the inhibitory herbicide or an analogous inhibitor suitable for use in a laboratory environment. Varying degrees of growth are recorded in different cultures. In certain cultures, fast-growing variant colonies arise that continue to grow even in the presence of normally inhibitory concentrations of inhibitor. The frequency with which such faster-growing variants occur can be increased by treatment with a chemical or physical mutagen before exposing the tissues or cells to the inhibitor. Putative tolerance-conferring alleles of the AIR synthetase gene are isolated and tested as described in the foregoing paragraphs. Those alleles identified as conferring herbicide tolerance may then be engineered for optimal expression and transformed into the plant. Alternatively, plants can be regenerated from the tissue or cell cultures containing these alleles.

Still another method involves mutagenesis of wild-type, herbicide sensitive plant AIR synthetase genes in bacteria or yeast, followed by culturing the microbe on medium that contains inhibitory concentrations of the inhibitor and then selecting those colonies that grow in the presence of the inhibitor. More specifically, a plant cDNA, such as the *Arabidopsis* cDNA encoding the AIR synthetase is cloned into a microbe that otherwise lacks the selected gene's activity. The transformed microbe is then subjected to *in vivo* mutagenesis or to *in vitro* mutagenesis by any of several chemical or enzymatic methods known in the art, e.g. sodium bisulfite (Shortle *et al.*, *Methods Enzymol.* 100:457-468 (1983); methoxylamine (Kadonaga *et al.*, *Nucleic Acids Res.* 13:1733-1745 (1985); oligonucleotide-directed saturation mutagenesis (Hutchinson *et al.*, *Proc. Natl. Acad. Sci. USA*, 83:710-714 (1986); or various polymerase misincorporation strategies (see, e.g. Shortle *et al.*, *Proc. Natl. Acad. Sci. USA*, 79:1588-1592 (1982); Shiraishi *et al.*, *Gene* 64:313-319 (1988); and Leung *et al.*, *Technique* 1:11-15 (1989). Colonies that grow in the presence of normally inhibitory concentrations of inhibitor are picked and purified by repeated restreaking. Their plasmids are purified and tested for the ability to confer tolerance to the inhibitor by retransforming them into the microbe lacking AIR synthetase gene activity. The DNA sequences of cDNA inserts from plasmids that pass this test are then determined.

Herbicide resistant AIR synthetase enzymes are also obtained using methods involving *in vitro* recombination, also called DNA shuffling. By DNA shuffling, mutations, preferably random mutations, are introduced in AIR synthetase genes. DNA shuffling also

leads to the recombination and rearrangement of sequences within an AIR synthetase genes or to recombination and exchange of sequences between two or more different of AIR synthetase genes. These methods allows for the production of millions of mutated AIR synthetase genes. The mutated genes, or shuffled genes, are screened for desirable properties, e.g. improved tolerance to herbicides and for mutations that provide broad spectrum tolerance to the different classes of inhibitor chemistry. Such-screens are well within the skills of a routineer in the art.

In a preferred embodiment, a mutagenized AIR synthetase gene is formed from at least one template AIR synthetase gene, wherein the template AIR synthetase gene has been cleaved into double-stranded random fragments of a desired size, and comprising the steps of adding to the resultant population of double-stranded random fragments one or more single or double-stranded oligonucleotides, wherein said oligonucleotides comprise an area of identity and an area of heterology to the double-stranded random fragments; denaturing the resultant mixture of double-stranded random fragments and oligonucleotides into single-stranded fragments; incubating the resultant population of single-stranded fragments with a polymerase under conditions which result in the annealing of said single-stranded fragments at said areas of identity to form pairs of annealed fragments, said areas of identity being sufficient for one member of a pair to prime replication of the other, thereby forming a mutagenized double-stranded polynucleotide; and repeating the second and third steps for at least two further cycles, wherein the resultant mixture in the second step of a further cycle includes the mutagenized double-stranded polynucleotide from the third step of the previous cycle, and the further cycle forms a further mutagenized double-stranded polynucleotide, wherein the mutagenized polynucleotide is a mutated AIR synthetase gene having enhanced tolerance to a herbicide which inhibits naturally occurring AIR synthetase activity. In a preferred embodiment, the concentration of a single species of double-stranded random fragment in the population of double-stranded random fragments is less than 1% by weight of the total DNA. In a further preferred embodiment, the template double-stranded polynucleotide comprises at least about 100 species of polynucleotides. In another preferred embodiment, the size of the double-stranded random fragments is from about 5 bp to 5 kb. In a further preferred embodiment, the fourth step of the method comprises repeating the second and the third steps for at least 10 cycles. Such method is described e.g. in Stemmer et al. (1994) Nature 370: 389-391, in US Patent 5,605,793 and in Cramer et al. (1998) Nature 391: 288-291, as well as in WO 97/20078, and these references are incorporated herein by reference.

In another preferred embodiment, any combination of two or more different AIR synthetase genes are mutagenized *in vitro* by a staggered extension process (StEP), as described e.g. in Zhao et al. (1998) Nature Biotechnology 16: 258-261. The two or more AIR synthetase genes are used as template for PCR amplification with the extension cycles of the PCR reaction preferably carried out at a lower temperature than the optimal polymerization temperature of the polymerase. For example, when a thermostable polymerase with an optimal temperature of approximately 72°C is used, the temperature for the extension reaction is desirably below 72°C, more desirably below 65°C, preferably below 60°C, more preferably the temperature for the extension reaction is 55°C. Additionally, the duration of the extension reaction of the PCR cycles is desirably shorter than usually carried out in the art, more desirably it is less than 30 seconds, preferably it is less than 15 seconds, more preferably the duration of the extension reaction is 5 seconds. Only a short DNA fragment is polymerized in each extension reaction, allowing template switch of the extension products between the starting DNA molecules after each cycle of denaturation and annealing, thereby generating diversity among the extension products. The optimal number of cycles in the PCR reaction depends on the length of the AIR synthetase coding regions to be mutagenized but desirably over 40 cycles, more desirably over 60 cycles, preferably over 80 cycles are used. Optimal extension conditions and the optimal number of PCR cycles for every combination of AIR synthetase genes are determined as described in using procedures well-known in the art. The other parameters for the PCR reaction are essentially the same as commonly used in the art. The primers for the amplification reaction are preferably designed to anneal to DNA sequences located outside of the coding sequence of the AIR synthetase genes, e.g. to DNA sequences of a vector comprising the AIR synthetase genes, whereby the different AIR synthetase genes used in the PCR reaction are preferably comprised in separate vectors. The primers desirably anneal to sequences located less than 500 bp away from the AIR synthetase coding sequences, preferably less than 200 bp away from the AIR synthetase coding sequences, more preferably less than 120 bp away from the AIR synthetase coding sequences. Preferably, the AIR synthetase coding sequences are surrounded by restriction sites, which are included in the DNA sequence amplified during the PCR reaction, thereby facilitating the cloning of the amplified products into a suitable vector.

In another preferred embodiment, fragments of AIR synthetase genes having cohesive ends are produced as described in WO 98/05765. The cohesive ends are produced by ligating a first oligonucleotide corresponding to a part of a AIR synthetase

gene to a second oligonucleotide not present in the gene or corresponding to a part of the gene not adjoining to the part of the gene corresponding to the first oligonucleotide, wherein the second oligonucleotide contains at least one ribonucleotide. A double-stranded DNA is produced using the first oligonucleotide as template and the second oligonucleotide as primer. The ribonucleotide is cleaved and removed. The nucleotide(s) located 5' to the ribonucleotide is also removed, resulting in double-stranded fragments having cohesive ends. Such fragments are randomly reassembled by ligation to obtain novel combinations of gene sequences.

Any AIR synthetase gene or any combination of AIR synthetase genes is used for *in vitro* recombination in the context of the present invention, for example, an AIR synthetase gene derived from a plant, such as, e.g. *Arabidopsis thaliana*, e.g. an AIR synthetase gene set forth in SEQ ID NO:1 or SEQ ID NO:3, an AIR synthetase gene from a bacteria, such as *Bacillus subtilis* (Ebbola and Zalkin (1987) J. Biol. Chem. 262: 8274-8287) or *E. coli* (Smith and Daum (1986) J. Biol. Chem. 261: 10632-10637), a human AIR synthetase gene (Aimi et al. (1990) Nucleic Acids Res. 18: 6665-6672), or an AIR synthetase gene from *Drosophila* (Henikoff et al. (1986) PNAS 83: 33-37), from chicken (Chen et al. (1990) PNAS 87: 3097-3101), and all incorporated herein by reference. Whole AIR synthetase genes or portions thereof are used in the context of the present invention. The library of mutated AIR synthetase genes obtained by the methods described above are cloned into appropriate expression vectors and the resulting vectors are transformed into an appropriate host, for example an algae like *Chlamydomonas*, a yeast or a bacteria. An appropriate host is preferably a host that otherwise lacks AIR synthetase gene activity, for example *E. coli* strain SØ6609/IKC (Schnorr et al. (1994) Plant Journal 6: 113-121). Host cells transformed with the vectors comprising the library of mutated AIR synthetase genes are cultured on medium that contains inhibitory concentrations of the inhibitor and those colonies that grow in the presence of the inhibitor are selected. Colonies that grow in the presence of normally inhibitory concentrations of inhibitor are picked and purified by repeated restreaking. Their plasmids are purified and the DNA sequences of cDNA inserts from plasmids that pass this test are then determined.

An assay for identifying a modified AIR synthetase gene that is tolerant to an inhibitor may be performed in the same manner as the assay to identify inhibitors of the AIR synthetase activity (Inhibitor Assay, above) with the following modifications: First, a mutant AIR synthetase is substituted in one of the reaction mixtures for the wild-type AIR synthetase of the inhibitor assay. Second, an inhibitor of wild-type enzyme is present in

both reaction mixtures. Third, mutated activity (activity in the presence of inhibitor and mutated enzyme) and unmutated activity (activity in the presence of inhibitor and wild-type enzyme) are compared to determine whether a significant increase in enzymatic activity is observed in the mutated activity when compared to the unmutated activity. Mutated activity is any measure of activity of the mutated enzyme while in the presence of a suitable substrate and the inhibitor. Unmutated activity is any measure of activity of the wild-type enzyme while in the presence of a suitable substrate and the inhibitor. A significant increase is defined as an increase in enzymatic activity that is larger than the margin of error inherent in the measurement technique, preferably an increase by about 2-fold or greater of the activity of the wild-type enzyme in the presence of the inhibitor, more preferably an increase by about 5-fold or greater, most preferably an increase by about 10-fold or greater.

In addition to being used to create herbicide-tolerant plants, genes encoding herbicide tolerant AIR synthetases can also be used as selectable markers in plant cell transformation methods. For example, plants, plant tissue, plant seeds, or plant cells transformed with a transgene can also be transformed with a gene encoding an altered AIR synthetase capable of being expressed by the plant. The transformed cells are transferred to medium containing an inhibitor of the enzyme in an amount sufficient to inhibit the survivability of plant cells not expressing the modified gene, wherein only the transformed cells will survive. The method is applicable to any plant cell capable of being transformed with a modified AIR synthetase-encoding gene, and can be used with any transgene of interest. Expression of the transgene and the modified gene can be driven by the same promoter functional in plant cells, or by separate promoters.

VII. Plant Transformation Technology

A wild-type or herbicide-tolerant form of the AIR synthetase gene can be incorporated in plant or bacterial cells using conventional recombinant DNA technology. Generally, this involves inserting a DNA molecule encoding the AIR synthetase into an expression system to which the DNA molecule is heterologous (i.e., not normally present) using standard cloning procedures known in the art. The vector contains the necessary elements for the transcription and translation of the inserted protein-coding sequences in a host cell containing the vector. A large number of vector systems known in the art can be used, such as plasmids, bacteriophage viruses and other modified viruses. The components of the expression system may also be modified to increase expression. For

example, truncated sequences, nucleotide substitutions or other modifications may be employed. Expression systems known in the art can be used to transform virtually any crop plant cell under suitable conditions. A transgene comprising a wild-type or herbicide-tolerant form of the AIR synthetase gene is preferably stably transformed and integrated into the genome of the host cells. In another preferred embodiment, the transgene comprising a wild-type or herbicide-tolerant form of the AIR synthetase gene located on a self-replicating vector. Examples of self-replicating vectors are viruses, in particular gemini viruses. Transformed cells can be regenerated into whole plants such that the chosen form of the AIR synthetase gene confers herbicide tolerance in the transgenic plants.

A. Requirements for Construction of Plant Expression Cassettes

Gene sequences intended for expression in transgenic plants are first assembled in expression cassettes behind a suitable promoter expressible in plants. The expression cassettes may also comprise any further sequences required or selected for the expression of the transgene. Such sequences include, but are not restricted to, transcription terminators, extraneous sequences to enhance expression such as introns, vital sequences, and sequences intended for the targeting of the gene product to specific organelles and cell compartments. These expression cassettes can then be easily transferred to the plant transformation vectors described *infra*. The following is a description of various components of typical expression cassettes.

1. Promoters

The selection of the promoter used in expression cassettes will determine the spatial and temporal expression pattern of the transgene in the transgenic plant. Selected promoters will express transgenes in specific cell types (such as leaf epidermal cells, mesophyll cells, root cortex cells) or in specific tissues or organs (roots, leaves or flowers, for example) and the selection will reflect the desired location of accumulation of the gene product. Alternatively, the selected promoter may drive expression of the gene under various inducing conditions. Promoters vary in their strength, i.e., ability to promote transcription. Depending upon the host cell system utilized, any one of a number of suitable promoters known in the art can be used. For example, for constitutive expression, the CaMV 35S promoter, the rice actin promoter, or the ubiquitin promoter may be used. For regulatable expression, the chemically inducible PR-1 promoter from tobacco or *Arabidopsis* may be used (see, e.g., U.S. Patent No. 5,689,044).

2. Transcriptional Terminators

A variety of transcriptional terminators are available for use in expression cassettes. These are responsible for the termination of transcription beyond the transgene and its correct polyadenylation. Appropriate transcriptional terminators are those that are known to function in plants and include the CaMV 35S terminator, the *tm1* terminator, the nopaline synthase terminator and the pea *rbcS* E9 terminator. These can be used in both monocotyledonous and dicotyledonous plants.

3. Sequences for the Enhancement or Regulation of Expression

Numerous sequences have been found to enhance gene expression from within the transcriptional unit and these sequences can be used in conjunction with the genes of this invention to increase their expression in transgenic plants. For example, various intron sequences such as introns of the maize *Adhl* gene have been shown to enhance expression, particularly in monocotyledonous cells. In addition, a number of non-translated leader sequences derived from viruses are also known to enhance expression, and these are particularly effective in dicotyledonous cells.

4. Coding Sequence Optimization

The coding sequence of the selected gene may be genetically engineered by altering the coding sequence for optimal expression in the crop species of interest. Methods for modifying coding sequences to achieve optimal expression in a particular crop species are well known (see, e.g. Perlak *et al.*, *Proc. Natl. Acad. Sci. USA* 88: 3324 (1991); and Koziel *et al.*, *Bio/technol.* 11: 194 (1993)).

5. Targeting of the Gene Product Within the Cell

Various mechanisms for targeting gene products are known to exist in plants and the sequences controlling the functioning of these mechanisms have been characterized in some detail. For example, the targeting of gene products to the chloroplast is controlled by a signal sequence found at the amino terminal end of various proteins which is cleaved during chloroplast import to yield the mature protein (e.g. Comai *et al.* *J. Biol. Chem.* 263: 15104-15109 (1988)). Other gene products are localized to other organelles such as the mitochondrion and the peroxisome (e.g. Unger *et al.* *Plant Molec. Biol.* 13: 411-418 (1989)).

The cDNAs encoding these products can also be manipulated to effect the targeting of heterologous gene products to these organelles. In addition, sequences have been characterized which cause the targeting of gene products to other cell compartments. Amino terminal sequences are responsible for targeting to the ER, the apoplast, and extracellular secretion from aleurone cells (Koehler & Ho, *Plant Cell* 2: 769-783 (1990)). Additionally, amino terminal sequences in conjunction with carboxy terminal sequences are responsible for vacuolar targeting of gene products (Shinshi *et al.* *Plant Molec. Biol.* 14: 357-368 (1990)). By the fusion of the appropriate targeting sequences described above to transgene sequences of interest it is possible to direct the transgene product to any organelle or cell compartment.

B. Construction of Plant Transformation Vectors

Numerous transformation vectors available for plant transformation are known to those of ordinary skill in the plant transformation arts, and the genes pertinent to this invention can be used in conjunction with any such vectors. The selection of vector will depend upon the preferred transformation technique and the target species for transformation. For certain target species, different antibiotic or herbicide selection markers may be preferred. Selection markers used routinely in transformation include the *nptII* gene, which confers resistance to kanamycin and related antibiotics (Messing & Vierra, *Gene* 19: 259-268 (1982); Bevan *et al.*, *Nature* 304:184-187 (1983)), the *bar* gene, which confers resistance to the herbicide phosphinothricin (White *et al.*, *Nucl. Acids Res* 18: 1062 (1990), Spencer *et al.* *Theor. Appl. Genet* 79: 625-631 (1990)), the *hph* gene, which confers resistance to the antibiotic hygromycin (Blochinger & Diggelmann, *Mol Cell Biol* 4: 2929-2931), and the *dhfr* gene, which confers resistance to methatrexate (Bourouis *et al.*, *EMBO J.* 2(7): 1099-1104 (1983)), and the EPSPS gene, which confers resistance to glyphosate (U.S. Patent Nos. 4,940,935 and 5,188,642).

1. Vectors Suitable for *Agrobacterium* Transformation

Many vectors are available for transformation using *Agrobacterium tumefaciens*. These typically carry at least one T-DNA border sequence and include vectors such as pBIN19 (Bevan, *Nucl. Acids Res.* (1984)) and pXYZ. Typical vectors suitable for *Agrobacterium* transformation include the binary vectors pCIB200 and pCIB2001, as well as

the binary vector pClB10 and hygromycin selection derivatives thereof. (See, for example, U.S. Patent No. 5,639,949).

2. Vectors Suitable for non-*Agrobacterium* Transformation

Transformation without the use of *Agrobacterium tumefaciens* circumvents the requirement for T-DNA sequences in the chosen transformation vector and consequently vectors lacking these sequences can be utilized in addition to vectors such as the ones described above which contain T-DNA sequences. Transformation techniques that do not rely on *Agrobacterium* include transformation via particle bombardment, protoplast uptake (e.g. PEG and electroporation) and microinjection. The choice of vector depends largely on the preferred selection for the species being transformed. Typical vectors suitable for non-*Agrobacterium* transformation include pClB3064, pSOG19, and pSOG35. (See, for example, U.S. Patent No. 5,639,949).

C. Transformation Techniques

Once the coding sequence of interest has been cloned into an expression system, it is transformed into a plant cell. Methods for transformation and regeneration of plants are well known in the art. For example, Ti plasmid vectors have been utilized for the delivery of foreign DNA, as well as direct DNA uptake, liposomes, electroporation, micro-injection, and microprojectiles. In addition, bacteria from the genus *Agrobacterium* can be utilized to transform plant cells.

Transformation techniques for dicotyledons are well known in the art and include *Agrobacterium*-based techniques and techniques that do not require *Agrobacterium*. Non-*Agrobacterium* techniques involve the uptake of exogenous genetic material directly by protoplasts or cells. This can be accomplished by PEG or electroporation mediated uptake, particle bombardment-mediated delivery, or microinjection. In each case the transformed cells are regenerated to whole plants using standard techniques known in the art.

Transformation of most monocotyledon species has now also become routine. Preferred techniques include direct gene transfer into protoplasts using PEG or electroporation techniques, particle bombardment into callus tissue, as well as *Agrobacterium*-mediated transformation.

VIII. Breeding

The wild-type or altered form of a AIR synthetase gene of the present invention can be utilized to confer herbicide tolerance to a wide variety of plant cells, including those of gymnosperms, monocots, and dicots. Although the gene can be inserted into any plant cell falling within these broad classes, it is particularly useful in crop plant cells, such as rice, wheat, barley, rye, corn, potato, carrot, sweet potato, sugar beet, bean, pea, chicory, lettuce, cabbage, cauliflower, broccoli, turnip, radish, spinach, asparagus, onion, garlic, eggplant, pepper, celery, carrot, squash, pumpkin, zucchini, cucumber, apple, pear, quince, melon, plum, cherry, peach, nectarine, apricot, strawberry, grape, raspberry, blackberry, pineapple, avocado, papaya, mango, banana, soybean, tobacco, tomato, sorghum and sugarcane.

The high-level expression of a wild-type AIR synthetase gene and/or the expression of herbicide-tolerant forms of a AIR synthetase gene conferring herbicide tolerance in plants, in combination with other characteristics important for production and quality, can be incorporated into plant lines through breeding approaches and techniques known in the art.

Where a herbicide tolerant AIR synthetase gene allele is obtained by direct selection in a crop plant or plant cell culture from which a crop plant can be regenerated, it is moved into commercial varieties using traditional breeding techniques to develop a herbicide tolerant crop without the need for genetically engineering the allele and transforming it into the plant.

The invention will be further described by reference to the following detailed examples. These examples are provided for purposes of illustration only, and are not intended to be limiting unless otherwise specified.

EXAMPLES

Standard recombinant DNA and molecular cloning techniques used here are well known in the art and are described by Sambrook, *et al.*, Molecular Cloning, eds., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (1989) and by T.J. Silhavy, M.L. Berman, and L.W. Enquist, Experiments with Gene Fusions, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1984) and by Ausubel, F.M. *et al.*, Current Protocols in Molecular Biology, pub. by Greene Publishing Assoc. and Wiley-Interscience (1987).

Example 1: Construction of a Vector Containing a GAL4 Binding Site/Minimal 35S CaMV Promoter Fused to Antisense AIR Synthetase

pAT71:

GAL4 binding sites and the minimal 35S promoter (-59 to +1) are excised from pGALLuc2 (Goff, *et al.*, (1991) *Genes & Development* 5: 298-309) as an *EcoRI-PstI* fragment and inserted into the respective sites of pBluescript, yielding pAT52. pAT66 is constructed with a three-way ligation between the *HindIII-PstI* fragment of pAT52, a *PstI-EcoRI* fragment of pClB1716 (contains a 35S untranslated leader, GUS gene, 35S terminator) and *HindIII-EcoRI* cut pUC18. The 35S leader of pAT66 is excised with *PstI-NcoI* and replaced with a PCR-generated 35S leader extending from +1 to +48 to yield pAT71.

pJG304:

Plasmid pBS SK+ (Stratagene, LaJolla, CA) is linearized with *SacI*, treated with mung bean nuclease to remove the *SacI* site, and re-ligated with T4 ligase to make pJG201. The 10XGAL4 consensus binding site/CaMV 35S minimal promoter/GUS gene/CaMV terminator cassette is removed from pAT71 with *KpnI* and cloned into the *KpnI* site of pJG201 to make pJG304.

pJG304 is partially digested with restriction endonuclease Asp718 to isolate a full-length linear fragment. This fragment is ligated with a molar excess of the 22 base oligonucleotide JG-L (5' GTACCTCGAG TCTAGACTCG AG 3', SEQ ID NO:5). Restriction analysis is used to identify a clone with this linker inserted 5' to the GAL4 DNA binding site, and this plasmid is designated pJG304DXhoI.

pDG3:

A fragment of the 5' phosphoribosyl-5-aminoimidazole (AIR) synthetase cDNA clone (Senecoff and Meagher (1993) Plant Physiology, 102: 387-399) is PCR-amplified from the *Arabidopsis thaliana* cDNA plasmid library pFL61 (Minet et al, (1992) Plant Journal, 2: 417-422) using the oligonucleotides AS-1 (5' GAT CGA GCT CGT TCT CTT CTG TGT CAT C 3', SEQ ID NO:6) and AS-2 (5' GAT CCC ATG GTC CCC AGG TAA AGA CGT C 3', SEQ ID NO:7).

The vector pJG304ΔXhoI is digested with *SacI* and *NcoI* to excise the GUS gene coding sequence. The AIR synthetase PCR fragment is digested with *SacI* and *NcoI* and ligated into pJG304ΔXhoI to make pDG3.

Example 2: Plant Transformation Vectors for AIR Synthetase Antisense Expression from the GAL4 Binding Site/CaMV Minimal 35S Promoter

pJG261:

Vector pGPTV (Becker, et al., (1992) Plant Molecular Biology 20: 1195-1197) is digested with *EcoRI* and *HindIII* to remove the nopaline synthase promoter/GUS cassette. Concurrently, the superlinker is excised from pSE380 (Invitrogen, San Diego, CA) with *EcoRI* and *HindIII* and cloned into the *EcoRI*/*HindIII* linearized pGPTV, to make pJG261.

PDG4:

pDG3 is cut with *XhoI* to excise the cassette containing the GAL4 DNA binding site/35S minimal promoter/antisense AIR synthetase/CaMV terminator fusion. This cassette is ligated into *XhoI* digested pJG261, such that transcription is divergent from that of the BAR selectable marker, producing pDG4.

Example 3: Production of GAL4 Binding Site/Minimal CaMV 35S Antisense AIR Synthetase Transgenic Plants

pDG4 is electro-transformed (Bio-Rad Laboratories, Hercules, CA) into *Agrobacterium tumefaciens* strain C58C1 (pMP90), and *Arabidopsis* plants (Ecotype Columbia) are transformed by infiltration (Bechtold et al, C. R. Acad. Sci. Paris, 316: 1188-

- 37 -

93 (1993). Seeds from the infiltrated plants are selected on germination medium (Murashige-Skoog salts at 4.3 g/liter, Mes at 0.5 g/liter, 1% sucrose, thiamine at 10 ug/liter, pyridoxine at 5 ug/liter, nicotinic acid at 5 ug/liter, myo-inositol at 1 mg/liter, pH 5.8) containing Basta at 15 mg/liter.

Example 4: Production of GAL4/C1 Transactivator Transgenic Plants

pSGZL1 is constructed by ligating the GAL4-C1 *EcoRI* fragment from pGALC1 (Goff, *et al.*, (1991) *Genes & Development*, 5: 298-309) into the *EcoRI* site of pIC20H. The GAL4-C1 fragment of pSGZL1 is excised with *BamHI-BglII* and inserted into the *BamHI* site of pCIB770 (Rothstein, *et al.*, (1987) *Gene* 53: 153-161) yielding pAT53.

Arabidopsis root explants are transformed with pAT53 as described in Valvekens, *et al.*, (1985) *PNAS USA* 85: 5536-5540. Transgenic plants with single site insertion and positive for GAL4/C1 expression are taken to homozygosity.

Example 5: Antisense Inhibition of AIR Synthetase Using a GAL4/C1 Transactivator and a GAL4 Binding Site/Minimal CaMV 35S Promoter

Fifteen transgenic plants containing the GAL4 binding site/minimal CaMV 35S promoter/antisense AIR synthetase construct are transplanted to soil and grown to maturity in the greenhouse. Flowers borne on the primary transformants are crossed to pollen from the homozygous GAL4/C1 transactivator line pAT53-103. F1 seeds are plated on germination medium and germination medium containing 15 mg/liter Basta. Seedlings from five F1 lines are transplanted to soil and grown to maturity in the greenhouse. Half of the seedlings from two F1 lines die while in soil. Half of the seedlings from three F1 lines are bleached and severely retarded in growth. These results show that the AIR synthetase gene is essential in plants.

Example 6: Expression of Recombinant Plant AIR Synthetase in *E. coli*

An *Arabidopsis thaliana* (Landsberg) cDNA library in the plasmid vector pFL61 (Minet *et al.*, *Plant J.*, 2:417-422 (1992)) is obtained and amplified. PCR primers to amplify protein coding sequence of *Arabidopsis* AIR synthetase are designed from a published

DNA sequence (Genbank accession L12457, Senecoff and Meagher, Plant Physiol., 102: 387-399 (1993)) and used to amplify the AIR synthetase coding sequence from the plasmid library with Pfu DNA polymerase (Stratagene). Sequencing of the PCR product reveals an error in the published DNA sequence resulting in the insertion of a cytosine base at the position corresponding to position 1027 in SEQ ID NO:1, resulting in an incorrect predicted protein. Several other changes such as a 9 bp (= 3 amino acid) insertion in the chloroplast transit peptide are observed but are probably due to variation between ecotypes. Primers are redesigned to correspond to the correct coding sequence. For the construct including the coding region of the AIR synthetase pre-protein, primers slp242 (5' CGC GGA TCC TCA CTA CTG ATA GCT TAC GCC TTC ACC 3', SEQ ID NO:8) and slp244 (5' TTG AAG CCA TGG AAG CTC GGA TTT TG 3', SEQ ID NO:9) are used, and for the construct including the coding region of the putative mature AIR synthetase, primers slp242 and slp243 (5' CGC ATG CCA TGG ATA AAG ATG ATG ACA CTG ATA GTC T 3', SEQ ID NO:10) are used. The coding regions of the pre-protein and of the putative mature protein are subcloned into the expression vector pET32a (Novagen) and both are transformed into *E.coli* BL21 DE3 pLysS (Novagen) by electroporation using the Biorad Gene Pulser and the manufacturer's conditions.

Example 7: Growth and Extraction of FGAM Synthetase

E.coli strain TX635/pJS113 (Schendel et al. (1989) Biochemistry 28, 2459-2471) is grown in Luria broth (LB) containing 50 µg/mL carbenicillin at 30°C in an incubator/shaker. When the cells reach an optical density of approximately 1 OD at 600 nm, an equal volume of LB carbenicillin at 56°C is added to heat-shock the cells. Subsequently, the cells are placed in an incubator-shaker and grown at 42°C. The cells are harvested at the end of log phase using low speed centrifugation. The centrifuge bottle is inverted and the media is allowed to drain. The cell pellet is resuspended with a small paintbrush in buffer A (50 mM EPPS, pH 7.5, 1mM EDTA, 2 mM DTT, 150 mM KCl, 10% glycerol) and then disrupted in a french pressure cell at 18,000 PSI. Following a high speed centrifugation to remove cell debris, the enzyme is precipitated with ammonium sulfate (40-60%) and the pellets stored at -80°C. The enzyme is resuspended in a small volume of Buffer A and applied to a Sephadex G-25 column for desalting into Buffer A. The activity is assayed as described below.

Example 8: Growth and Extraction of AIR Synthetase

E.coli strain pJS24/Tx393 (Schrimsher et al. (1986) Biochemistry 25, 4366-4371) containing multiple gene copies of the native AIR synthetase is grown in LB broth containing 50 µg/mL of carbenicillin at 37°C in an incubator-shaker. The cells are harvested at the end of the log phase of growth and pelleted in a centrifuge at low speed, the growth media is discarded and the centrifuge bottle is inverted and allowed to drain. The cells are resuspended in buffer A with a small paintbrush and disrupted in a French Pressure Cell at approximately 18,000 PSI. Following a high speed centrifugation to pellet cell debris, the supernatant is precipitated with ammonium sulfate and stored at -80°C.

Example 9: AIR Synthetase Activity Assay

The AIR synthetase activity assay is essentially derived from Schrimsher et al. (1986) Biochemistry 25, 4356-4365. The reaction volumes are preferably the ones described below, but can be varied depending on the experimental requirements. 0.2-1.0 x 10⁻⁴ unit of an enzyme having AIR synthetase activity (one unit of activity is defined as the amount of enzyme required to produce 1 mmol/min of product) and 0.1 mM 5'-phosphoribosyl-N-formylglycinamide (FGAM) are mixed in a final volume of 96 µl 50 mM HEPES (pH 7.4-8.1, but preferably 7.7), 20 mM MgCl₂, 150 mM KCl and 0.01-10 mM, but preferably 2.0 mM ATP. The production of AIR is determined preferably according to Bratton and Marshall (J. Biol. Chem. (1939) 128, 537-550) by adding 32 µl of 1.33 M potassium phosphate in 20% (w/v) trichloroacetic acid (pH 1.4). The mixture is centrifuged to remove precipitated protein and 32 µl of 0.1% (w/v) sodium nitrite is added. After 3 min., 32 µl of 0.5% (w/v) ammonium sulfamate is added and, after an additional minute, 8 µl of 25% N-(1-naphthyl)ethylenediamine dihydrochloride is added. The absorbance is measured at 530 nm after 10 min.

Alternatively, ADP formation is quantitated by a coupled reaction procedure. In this case, 3.5 units of pyruvate kinase, 4.7 units of lactate dehydrogenase, 1.0 mM phosphoenolpyruvate and 0.2 mM NADH are added and absorbance is measured at 340 nm.

Example 10: Coupled FGAM Synthetase and AIR Synthetase Enzyme Assays

A. FGAM synthetase assay

The conversion of FGAR to FGAM is followed by detecting the concomittant formation of ADP. The ADP formation is followed utilizing the enzymes pyruvate kinase, and lactate dehydrogenase (reagent enzymes) and detecting the conversion of NADH to NAD⁺ in the presence of phosphoenolpyruvate (PEP). This is monitored at 340 nm. Pyruvate kinase and PEP facilitate the regeneration of ATP from ADP. ATP is a required substrate for both FGAM synthetase and AIR synthetase. The assay buffer is buffer A with the addition of 20 mM MgCl₂.

B. AIR synthase assay

To assay AIR synthase it is necessary to provide the substrate FGAM. The FGAM is provided by the conversion of FGAR to FGAM in the same reaction mixture. If NADH is added the conversion can be followed utilizing the FGAM synthetase assay. When the FGAR-FGAM conversion proceeds sufficiently (approximately 50µM) then AIR synthetase is added. Adding the AIR synthetase after the production of FGAM insures that the initial concentration of FGAM is constant in all reaction wells. The AIR synthetase is assayed by the method of Bratton and Marshall (J. Biol. Chem. (1939) 128, 537-550). After a sufficient time for AIR production (typically 15 minutes) the enzyme reaction is stopped with TCA. The AIR is derivatized with sodium nitrite and the nitrite is subsequently neutralized with ammonium sulfamate. The color is developed with the addition of N-(1-naphthyl)ethylene-diamine dihydrochloride (NEDD). After 10 minutes the color is monitored at 530 nm.

C. Assay Protocols

The assays are carried out in the same way independent of the original source of the enzymes. The assays are performed in 300 µL 96 well microtiter plates. The total assay reaction volume is 200 µL. Substrates (except FGAR) are mixed in a ratio such that the final concentrations (in the microtiterplate) are as follows: L-glutamine (600 µM), ATP (600 µM), PEP (1 mM), and NADH (200 µM). A mixture of substrates at 10X concentration can be pipetted at 20 µL/well. The reagent enymes and FGAM synthetase can also be mixed to be added simultaneously. The suggested amounts of the ADP detecting/regeneration mix is 0.7 units pyruvate kinase and 0.97 units lactate dehydrogenase per reaction. This should be used as a guideline and the amounts of enzyme adjusted empirically. The FGAR (200

μM) should be added after a two minute incubation period. After the FGAM synthase reaction proceeds to completion at a rate of approximately $10 \mu\text{M}/\text{minute}$ (this is within 10-15 minutes), the AIR synthetase is added. After an interval (determined by the activity of the AIR synthetase) the reaction is stopped with $66 \mu\text{L}$ of 20% TCA in $1\text{M K}_3\text{PO}_4$. The plate is spun in a centrifuge to pellet the precipitated protein, then the supernatant is transferred to a separate microtiterplate for color development and reading. $1.2 \mu\text{L}$ of 10% sodium nitrite is added and after 3 minutes $1.2 \mu\text{L}$ of 50% ammonium sulfamate is added (neutralizes excess nitrite). One minute later, $8.3 \mu\text{L}$ of 1% NEDD are added and after 5 minutes, the plate is read at 530 nm using a microtiter plate reading UV/VIS spectrophotometer. AICAR is used as a standard since AIR is not available for that purpose. Based on AICAR a reasonable detection limit (3-fold OD over background) of $10 \mu\text{M}$ is easily attainable.

L-Glutamine, ATP, sodium nitrite, ammonium sulfamate, and NEDD, are available from Sigma Chemicals. FGAR is synthesized by the methods of Chen and Henderson (Can. J. Chemistry (1970) 48: 2306-2309) or Carrington et al. (J.Chem. Soc. (1968) 6864).

Example 11: In vitro Recombination of AIR Synthetase Genes by DNA Shuffling

The *A. thaliana* AIR synthetase gene encoding the pre-protein is amplified by PCR as described in example 6. The resulting DNA fragment is digested by DNase-I treatment essentially as described (Stemmer et al. (1994) PNAS 91: 10747-10751) and the PCR primers are removed from the reaction mixture. A PCR reaction is carried out without primers and is followed by a PCR reaction with the primers, both as described (Stemmer et al. (1994) PNAS 91: 10747-10751). The resulting DNA fragments are cloned into pTRC99a (Pharmacia, Cat no: 27-5007-01) and transformed into *E.coli* strain SØ6609/IKC (Schnorr et al. (1994) Plant Journal 6: 113-121) by electroporation using the Biorad Gene Pulser and the manufacturer's conditions. The transformed bacteria are grown on medium that contains inhibitory concentrations of the inhibitor and those colonies that grow in the presence of the inhibitor are selected. Colonies that grow in the presence of normally inhibitory concentrations of inhibitor are picked and purified by repeated restreaking. Their plasmids are purified and the DNA sequences of cDNA inserts from plasmids that pass this test are then determined.

In a similar reaction, PCR-amplified DNA fragments comprising the *A. thaliana* AIR synthetase gene encoding the pre-protein and PCR-amplified DNA fragments comprising

the *E.coli* purM gene are recombined *in vitro* and resulting variants with improved tolerance to the inhibitor are recovered as described above.

Example 12: In vitro Recombination of AIR Synthetase Genes by Staggered Extension Process

The *A. thaliana* AIR synthetase gene encoding the mature protein and the *E.coli* purM gene are each cloned into the polylinker of a pBluescript vector. A PCR reaction is carried out essentially as described (Zhao et al. (1998) Nature Biotechnology 16: 258-261) using the "reverse primer" and the "M13 20 primer" (Stratagene Catalog). Amplified PCR fragments are digested with appropriate restriction enzymes and cloned into pTRC99a and mutated AIR synthetase genes are screened as described in example 11.

The above disclosed embodiments are illustrative. This disclosure of the invention will place one skilled in the art in possession of many variations of the invention. All such obvious and foreseeable variations are intended to be encompassed by the appended claims.

- 43 -

What Is Claimed Is:

1. An isolated enzyme comprising an amino acid sequence that is identical or substantially similar to SEQ ID NO:2 or SEQ ID NO:4, wherein said enzyme has 5'-phosphoribosyl-5-aminoimidazole (AIR) synthetase activity.
2. An isolated enzyme according to claim 1, wherein said amino acid sequence is derived from a plant.
3. An isolated enzyme according to claim 1, wherein said amino acid sequence is SEQ ID NO:2.
4. An isolated enzyme according to claim 1, wherein said amino acid sequence is SEQ ID NO:4.
5. An isolated nucleic acid molecule comprising a nucleotide sequence that encodes the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4.
6. An isolated nucleic acid molecule according to claim 5, wherein said nucleotide sequence is SEQ ID NO:1 or SEQ ID NO:3.
7. An isolated nucleic acid molecule according to claim 5, wherein said nucleotide sequence is deposited in *E. coli* strain DH5apASM designated as NRRL B-21976.
8. A chimeric gene comprising a heterologous promoter sequence operatively linked to the nucleic acid molecule of claim 5.
9. A recombinant vector comprising the chimeric gene of claim 8.
10. A host cell comprising the chimeric gene of claim 8.
11. A host cell according to claim 10, which is a bacterial cell.

- 44 -

12. A host cell according to claim 10, which is a yeast cell.

13. A host cell according to claim 10, which is a plant cell.

14. A plant comprising the plant cell of claim 13.

15. Seed from the plant of claim 14.

16. A method of identifying chemicals having the ability to inhibit plant growth or viability, comprising:

- (a) combining an enzyme having AIR synthetase activity in a first reaction mixture with a substrate of AIR synthetase under conditions in which the enzyme is capable of catalyzing the synthesis of AIR;
- (b) combining a chemical to be tested and the enzyme in a second reaction mixture with a substrate of AIR synthetase under the same conditions and for the same period of time as in the first reaction mixture; and
- (c) determining and comparing the activity of the enzyme in the first and second reaction mixtures;

wherein less enzyme activity in the second reaction mixture than in the first reaction mixture indicates that the chemical of (b) has the ability to inhibit plant growth or viability.

17. A method of identifying chemicals having the ability to inhibit plant growth or viability, comprising:

- (a) combining an enzyme having 5'-phosphoribosyl-N-formylglycinamide (FGAM) synthetase activity and an enzyme having AIR synthetase activity in a first reaction mixture with a substrate of FGAM synthetase under conditions in which the enzymes are capable of catalyzing the coupled synthesis of AIR;
- (b) combining a chemical to be tested and the enzymes in a second reaction mixture with a substrate of FGAM synthetase under the same conditions and the same period of time as in the first reaction mixture; and
- (c) determining and comparing the activity of the enzyme having AIR synthetase activity in the first and second reaction mixtures;

wherein less enzyme activity in the second reaction mixture than in the first reaction mixture indicates that the chemical of (b) has the ability to inhibit plant growth or viability.

18. A method for identifying chemicals having herbicidal activity that inhibit AIR synthetase activity in plants, comprising:

- (a) obtaining transgenic plants, plant tissue, plant seeds or plant cells comprising an isolated nucleotide sequence encoding an enzyme having AIR synthetase activity and capable of overexpressing an enzymatically active AIR synthetase;
- (b) applying a chemical to be tested to the transgenic plants, plant cells, tissues or parts and to the isogenic non-transformed plants, plant cells, tissues or parts;
- (c) determining the growth or viability of the transgenic and non-transformed plants, plant cells, tissues after application of the chemical; and
- (d) comparing the growth or viability of the transgenic and non-transformed plants, plant cells, tissues after application of the chemical;

wherein suppression of the growth or viability of the non-transgenic plants, plant cells, tissues or parts, without significantly suppressing the growth or viability of the isogenic transgenic plants, plant cells, tissues or parts indicates that the chemical of (b) has herbicidal activity that inhibits AIR synthetase activity in plants.

19. A method according to claim 16, wherein the substrate is 5'-phosphoribosyl-N-formylglycinamide (FGAM).

20. A method according to claim 16, wherein the substrate is b-FGAM.

21. A method according to claim 17, wherein the substrate is 5'-phosphoribosyl-N-formylglycinamide (FGAR).

22. A method according to claim 17, wherein the substrate is b-FGAR.

23. A method according to any of claims 16-18, wherein the enzyme having AIR synthetase activity is derived from a plant.

24. A method according to any of claims 16-18, wherein the enzyme having AIR synthetase activity comprises an amino acid sequence identical or substantially similar to the amino acid sequence set forth in SEQ ID NO:2 or SEQ ID NO:4.

- 46 -

25. A method according to any of claims 16-18, wherein the enzyme having AIR synthetase activity is derived from *E. coli*.

26. A method according to any of claims 16-18, wherein the activity of the enzyme is determined by measuring the AIR produced in the reaction mixture.

27. A method for suppressing the growth of undesired vegetation, comprising the step of applying to the undesired vegetation a chemical identified by the method of any of claims 16-18.

28. A transgenic plant, plant cell, plant seed, or plant tissue comprising a nucleotide sequence encoding an enzyme having AIR synthetase activity, wherein the nucleotide sequence confers upon said transgenic plant, plant cell, plant seed, or plant tissue tolerance to a chemical identified by the method of any of claims 16-18 in amounts that normally inhibits AIR synthetase activity in a wild-type plant.

29. A plant made by a process comprising transforming the plant or a parent of the plant with an isolated DNA molecule comprising a nucleotide sequence encoding an enzyme having AIR synthetase activity and capable of expressing the nucleotide sequence in the plant so as to render the plant tolerant to a chemical identified by the method of any of claims 16-18.

30. A method for selectively suppressing the growth of weeds in a field containing a crop of planted crop seeds or plants, comprising:

(a) planting herbicide tolerant crops or crop seeds, which are plants or plant seeds transformed with an isolated DNA molecule comprising a nucleotide sequence having AIR synthetase activity, wherein said nucleotide sequence is expressible in said plant or plant seed; and

(b) applying to the crops or crop seeds and the weeds in the field a herbicide in amounts that inhibit naturally occurring AIR synthetase activity, wherein the herbicide suppresses the growth of the weeds without significantly suppressing the growth of the crops.

31. A method for forming a mutagenized DNA molecule encoding an enzyme having AIR synthetase activity from a template DNA molecule encoding an enzyme having AIR synthetase activity, wherein said template DNA molecule has been cleaved into double-stranded-random fragments, comprising the steps of:

- (a) adding to the resultant population of double-stranded-random fragments at least one single-stranded or double-stranded oligonucleotide, wherein said oligonucleotide comprises an area of identity and an area of heterology to the template DNA molecule;
- (b) denaturing the resultant mixture of double-stranded-random fragments and oligonucleotides into single-stranded molecules;
- (c) incubating the resultant population of single-stranded molecules with a polymerase under conditions which result in the annealing of said single-stranded molecules at said areas of identity to form pairs of annealed fragments, said areas of identity being sufficient for one member of a pair to prime replication of the other, thereby forming a mutagenized double-stranded polynucleotide;
- (d) repeating the second and third steps for at least two further cycles, wherein the resultant mixture in the second step of a further cycle includes the mutagenized double-stranded polynucleotide from the third step of the previous cycle, and the further cycle forms a further mutagenized double-stranded polynucleotide;

wherein the mutagenized double-stranded polynucleotide encodes an AIR synthetase enzyme having enhanced tolerance to a herbicide which inhibits the AIR synthetase activity encoded by the template DNA molecule.

32. A method for forming a mutagenized DNA molecule encoding an enzyme having AIR synthetase activity from at least two non-identical template DNA molecules encoding enzymes having AIR synthetase activity, comprising the steps of:

- (a) adding to the template DNA molecules at least one oligonucleotide comprising an area of identity to each of the template DNA molecule;
- (b) denaturing the resultant mixture into single-stranded molecules;
- (c) incubating the resultant population of single-stranded molecules with a polymerase under conditions which result in the annealing of the oligonucleotides to the template DNA molecules, wherein the conditions for

polymerization by the polymerase are such that polymerization products corresponding to a portion of the template DNA molecules are obtained;

- (d) repeating the second and third steps for at least two further cycles, wherein the extension products obtained in the third step are able to switch template DNA molecule for polymerization in the next cycle, thereby forming a mutagenized double-stranded polynucleotide comprising sequences derived from different template DNA molecules;

wherein the mutagenized double-stranded polynucleotide encodes an AIR synthetase enzyme having enhanced tolerance to a herbicide which inhibits the AIR synthetase activity encoded by the template DNA molecules.

33. A mutagenized DNA molecule encoding an enzyme having AIR synthetase activity obtained by the method of claim 31, wherein said mutagenized DNA molecule encodes an AIR synthetase enzyme having enhanced tolerance to a herbicide which inhibits the AIR synthetase activity encoded by said template DNA molecule.

34. A mutagenized DNA molecule encoding an enzyme having AIR synthetase activity obtained by the method of claim 32, wherein said mutagenized DNA molecule encodes an AIR synthetase enzyme having enhanced tolerance to a herbicide which inhibits the AIR synthetase activity encoded by said template DNA molecule.

35. The method of claim 31 or claim 32, wherein at least one template DNA molecule is derived from a eukaryote.

36. The method of claim 35, wherein said eukaryote is a plant.

37. The method of claim 36, wherein said plant is *Arabidopsis thaliana*.

38. The method of claim 37, wherein said species of template DNA molecule is identical or substantially similar to the SEQ ID NO:1 or SEQ ID NO:3.

39. The method of claim 31 or claim 32, wherein one template DNA molecule is derived from a prokaryote.

- 1 -

SEQUENCE LISTING

<110> Novartis AG

<120> METHODS TO SCREEN HERBICIDAL COMPOUNDS AND USES THEREOF

<130> PH/5-30552/A/CGC1999

<140>

<141>

<150> US 09/103,895

<151> 1998-06-24

<160> 10

<170> PatentIn Ver. 2.0

<210> 1

<211> 1172

<212> DNA

<213> Arabidopsis thaliana

<220>

<221> CDS

<222> (3)..(1160)

<223> AIR synthetase cDNA

<400> 1

```

cc atg gaa gct cgg att ttg cag tct tct tct tcc tgt tat tcg tct      47
   Met Glu Ala Arg Ile Leu Gln Ser Ser Ser Ser Cys Tyr Ser Ser
     1             5             10             15

ctt tac act gtc aat cga tcc cgg ttc tct tct ccg aaa cct ttc tcc      95
Leu Tyr Thr Val Asn Arg Ser Arg Phe Ser Ser Pro Lys Pro Phe Ser
           20             25             30

gtc agc ttt gct cag acg acg aga aca agg act cgt gta tta tcc atg      143
Val Ser Phe Ala Gln Thr Thr Arg Thr Arg Thr Arg Val Leu Ser Met
           35             40             45

tcg aag aaa gat ggt cgc act gat aaa gat gat gac act gat agt ctc      191
Ser Lys Lys Asp Gly Arg Thr Asp Lys Asp Asp Asp Thr Asp Ser Leu
           50             55             60

aat tac aaa gat tct ggt gtt gat atc gat gct ggt gct gag ctt gtt      239
Asn Tyr Lys Asp Ser Gly Val Asp Ile Asp Ala Gly Ala Glu Leu Val
           65             70             75

aaa cga atc gca aag atg gct cct gga att ggt gga ttt ggt ggt ctc      287
Lys Arg Ile Ala Lys Met Ala Pro Gly Ile Gly Gly Phe Gly Gly Leu
           80             85             90             95

ttt cca tta ggt gat agt tat ctt gta gct ggt acg gat ggt gta ggg      335
Phe Pro Leu Gly Asp Ser Tyr Leu Val Ala Gly Thr Asp Gly Val Gly
           100            105            110

act aaa ttg aaa ttg gca ttt gaa act gga att cat gac acc att gga      383
Thr Lys Leu Lys Leu Ala Phe Glu Thr Gly Ile His Asp Thr Ile Gly
           115            120            125

atc gac ttg gtt gct atg agt gtg aat gat att att act tct ggt gca      431
Ile Asp Leu Val Ala Met Ser Val Asn Asp Ile Ile Thr Ser Gly Ala
           130            135            140

```

- 2 -

aag cct ctg ttt ttc ctt gat tac ttt gct act agt cgt ctt gat gta	479
Lys Pro Leu Phe Phe Leu Asp Tyr Phe Ala Thr Ser Arg Leu Asp Val	
145 150 155	
gac ctt gct gaa aag gtc att aaa ggg att gtt gaa ggt tgt cgg caa	527
Asp Leu Ala Glu Lys Val Ile Lys Gly Ile Val Glu Gly Cys Arg Gln	
160 165 170 175	
tcg gaa tgt gct ctc tta ggg gga gag act gca gag atg cct gac ttt	575
Ser Glu Cys Ala Leu Leu Gly Gly Glu Thr Ala Glu Met Pro Asp Phe	
180 185 190	
tat gca gag ggc gag tac gat cta agt ggg ttt gca gta ggc ata gta	623
Tyr Ala Glu Gly Glu Tyr Asp Leu Ser Gly Phe Ala Val Gly Ile Val	
195 200 205	
aag aaa act tca gtt atc aac gga aaa aac att gtg gcc ggt gat gtt	671
Lys Lys Thr Ser Val Ile Asn Gly Lys Asn Ile Val Ala Gly Asp Val	
210 215 220	
ctt att ggc ctc ccg tct agt ggt gtt cat tcc aat ggt ttt tct cta	719
Leu Ile Gly Leu Pro Ser Ser Gly Val His Ser Asn Gly Phe Ser Leu	
225 230 235	
gta aga agg gta ttg gct cga agc aat ctt tcg ctg aat gat gcg ctt	767
Val Arg Arg Val Leu Ala Arg Ser Asn Leu Ser Leu Asn Asp Ala Leu	
240 245 250 255	
cca ggt gga tca agt acc ctt ggt gat gct cta atg gca ccc act gtc	815
Pro Gly Gly Ser Ser Thr Leu Gly Asp Ala Leu Met Ala Pro Thr Val	
260 265 270	
att tac gtg aaa cag gta ctt gat atg ata gaa aaa gga gga gtg aaa	863
Ile Tyr Val Lys Gln Val Leu Asp Met Ile Glu Lys Gly Gly Val Lys	
275 280 285	
ggt tta gct cat atc aca ggc gga ggt ttc aca gac aac att ccc cga	911
Gly Leu Ala His Ile Thr Gly Gly Gly Phe Thr Asp Asn Ile Pro Arg	
290 295 300	
gtc ttc ccg gac ggt ttg ggt gct gtt att cac acc gat act tgg gaa	959
Val Phe Pro Asp Gly Leu Gly Ala Val Ile His Thr Asp Thr Trp Glu	
305 310 315	
ctt cca ccg ttg ttc aag tgg att caa cag act ggg aga ata gaa gac	1007
Leu Pro Pro Leu Phe Lys Trp Ile Gln Gln Thr Gly Arg Ile Glu Asp	
320 325 330 335	
agt gag atg aga agg acg ttt aac ctg ggg ata ggg atg gtt atg gtg	1055
Ser Glu Met Arg Arg Thr Phe Asn Leu Gly Ile Gly Met Val Met Val	
340 345 350	
gtt agt cca gag gca gct tca cga ata cta gaa gaa gtc aag aat gga	1103
Val Ser Pro Glu Ala Ala Ser Arg Ile Leu Glu Glu Val Lys Asn Gly	
355 360 365	
gac tat gtt gcg tat cgc gta gga gag gtt gtc aac ggt gaa ggc gta	1151
Asp Tyr Val Ala Tyr Arg Val Gly Glu Val Val Asn Gly Glu Gly Val	
370 375 380	
agc tat cag tagtgaggat cc	1172
Ser Tyr Gln	
385	

- 3 -

<210> 2
 <211> 386
 <212> PRT
 <213> Arabidopsis thaliana

<400> 2
 Met Glu Ala Arg Ile Leu Gln Ser Ser Ser Ser Cys Tyr Ser Ser Leu
 1 5 10 15
 Tyr Thr Val Asn Arg Ser Arg Phe Ser Ser Pro Lys Pro Phe Ser Val
 20 25 30
 Ser Phe Ala Gln Thr Thr Arg Thr Arg Thr Arg Val Leu Ser Met Ser
 35 40 45
 Lys Lys Asp Gly Arg Thr Asp Lys Asp Asp Asp Thr Asp Ser Leu Asn
 50 55 60
 Tyr Lys Asp Ser Gly Val Asp Ile Asp Ala Gly Ala Glu Leu Val Lys
 65 70 75 80
 Arg Ile Ala Lys Met Ala Pro Gly Ile Gly Gly Phe Gly Gly Leu Phe
 85 90 95
 Pro Leu Gly Asp Ser Tyr Leu Val Ala Gly Thr Asp Gly Val Gly Thr
 100 105 110
 Lys Leu Lys Leu Ala Phe Glu Thr Gly Ile His Asp Thr Ile Gly Ile
 115 120 125
 Asp Leu Val Ala Met Ser Val Asn Asp Ile Ile Thr Ser Gly Ala Lys
 130 135 140
 Pro Leu Phe Phe Leu Asp Tyr Phe Ala Thr Ser Arg Leu Asp Val Asp
 145 150 155 160
 Leu Ala Glu Lys Val Ile Lys Gly Ile Val Glu Gly Cys Arg Gln Ser
 165 170 175
 Glu Cys Ala Leu Leu Gly Gly Glu Thr Ala Glu Met Pro Asp Phe Tyr
 180 185 190
 Ala Glu Gly Glu Tyr Asp Leu Ser Gly Phe Ala Val Gly Ile Val Lys
 195 200 205
 Lys Thr Ser Val Ile Asn Gly Lys Asn Ile Val Ala Gly Asp Val Leu
 210 215 220
 Ile Gly Leu Pro Ser Ser Gly Val His Ser Asn Gly Phe Ser Leu Val
 225 230 235 240
 Arg Arg Val Leu Ala Arg Ser Asn Leu Ser Leu Asn Asp Ala Leu Pro
 245 250 255
 Gly Gly Ser Ser Thr Leu Gly Asp Ala Leu Met Ala Pro Thr Val Ile
 260 265 270
 Tyr Val Lys Gln Val Leu Asp Met Ile Glu Lys Gly Gly Val Lys Gly
 275 280 285
 Leu Ala His Ile Thr Gly Gly Gly Phe Thr Asp Asn Ile Pro Arg Val
 290 295 300
 Phe Pro Asp Gly Leu Gly Ala Val Ile His Thr Asp Thr Trp Glu Leu
 305 310 315 320

- 4 -

Pro Pro Leu Phe Lys Trp Ile Gln Gln Thr Gly Arg Ile Glu Asp Ser
 325 330 335
 Glu Met Arg Arg Thr Phe Asn Leu Gly Ile Gly Met Val Met Val Val
 340 345 350
 Ser Pro Glu Ala Ala Ser Arg Ile Leu Glu Glu Val Lys Asn Gly Asp
 355 360 365
 Tyr Val Ala Tyr Arg Val Gly Glu Val Val Asn Gly Glu Gly Val Ser
 370 375 380
 Tyr Gln
 385

<210> 3
 <211> 1013
 <212> DNA
 <213> Arabidopsis thaliana

<220>
 <221> mat_peptide
 <222> (3)..(1001)
 <223> coding sequence of AIR synthetase putative mature
 peptide

<220>
 <221> CDS
 <222> (3)..(1001)

<400> 3
 cc atg gat aaa gat gat gac act gat agt ctc aat tac aaa gat tct 47
 Met Asp Lys Asp Asp Asp Thr Asp Ser Leu Asn Tyr Lys Asp Ser
 1 5 10 15
 ggt gtt gat atc gat gct ggt gct gag ctt gtt aaa cga atc gca aag 95
 Gly Val Asp Ile Asp Ala Gly Ala Glu Leu Val Lys Arg Ile Ala Lys
 20 25 30
 atg gct cct gga att ggt gga ttt ggt ggt ctc ttt cca tta ggt gat 143
 Met Ala Pro Gly Ile Gly Gly Phe Gly Gly Leu Phe Pro Leu Gly Asp
 35 40 45
 agt tat ctt gta gct ggt acg gat ggt gta ggg act aaa ttg aaa ttg 191
 Ser Tyr Leu Val Ala Gly Thr Asp Gly Val Gly Thr Lys Leu Lys Leu
 50 55 60
 gca ttt gaa act gga att cat gac acc att gga atc gac ttg gtt gct 239
 Ala Phe Glu Thr Gly Ile His Asp Thr Ile Gly Ile Asp Leu Val Ala
 65 70 75
 atg agt gtg aat gat att att act tct ggt gca aag cct ctg ttt ttc 287
 Met Ser Val Asn Asp Ile Ile Thr Ser Gly Ala Lys Pro Leu Phe Phe
 80 85 90 95
 ctt gat tac ttt gct act agt cgt ctt gat gta gac ctt gct gaa aag 335
 Leu Asp Tyr Phe Ala Thr Ser Arg Leu Asp Val Asp Leu Ala Glu Lys
 100 105 110
 gtc att aaa ggg att gtt gaa ggt tgt cgg caa tcg gaa tgt gct ctc 383
 Val Ile Lys Gly Ile Val Glu Gly Cys Arg Gln Ser Glu Cys Ala Leu
 115 120 125

- 5 -

tta ggg gga gag act gca gag atg cct gac ttt tat gca gag ggc gag 431
 Leu Gly Gly Glu Thr Ala Glu Met Pro Asp Phe Tyr Ala Glu Gly Glu
 130 135 140

tac gat cta agt ggg ttt gca gta ggc ata gta aag aaa act tca gtt 479
 Tyr Asp Leu Ser Gly Phe Ala Val Gly Ile Val Lys Lys Thr Ser Val
 145 150 155

atc aac gga aaa aac att gtg gcc ggt gat gtt ctt att ggc ctc ccg 527
 Ile Asn Gly Lys Asn Ile Val Ala Gly Asp Val Leu Ile Gly Leu Pro
 160 165 170 175

tct agt ggt gtt cat tcc aat ggt ttt tct cta gta aga agg gta ttg 575
 Ser Ser Gly Val His Ser Asn Gly Phe Ser Leu Val Arg Arg Val Leu
 180 185 190

gct cga agc aat ctt tcg ctg aat gat gcg ctt cca ggt gga tca agt 623
 Ala Arg Ser Asn Leu Ser Leu Asn Asp Ala Leu Pro Gly Gly Ser Ser
 195 200 205

acc ctt ggt gat gct cta atg gca ccc act gtc att tac gtg aaa cag 671
 Thr Leu Gly Asp Ala Leu Met Ala Pro Thr Val Ile Tyr Val Lys Gln
 210 215 220

gta ctt gat atg ata gaa aaa gga gga gtg aaa ggt tta gct cat atc 719
 Val Leu Asp Met Ile Glu Lys Gly Gly Val Lys Gly Leu Ala His Ile
 225 230 235

aca ggc gga ggt ttc aca gac aac att ccc cga gtc ttc ccg gac ggt 767
 Thr Gly Gly Gly Phe Thr Asp Asn Ile Pro Arg Val Phe Pro Asp Gly
 240 245 250 255

ttg ggt gct gtt att cac acc gat act tgg gaa ctt cca ccg ttg ttc 815
 Leu Gly Ala Val Ile His Thr Asp Thr Trp Glu Leu Pro Pro Leu Phe
 260 265 270

aag tgg att caa cag act ggg aga ata gaa gac agt gag atg aga agg 863
 Lys Trp Ile Gln Gln Thr Gly Arg Ile Glu Asp Ser Glu Met Arg Arg
 275 280 285

acg ttt aac ctg ggg ata ggg atg gtt atg gtg gtt agt cca gag gca 911
 Thr Phe Asn Leu Gly Ile Gly Met Val Met Val Val Ser Pro Glu Ala
 290 295 300

gct tca cga ata cta gaa gaa gtc aag aat gga gac tat gtt gcg tat 959
 Ala Ser Arg Ile Leu Glu Glu Val Lys Asn Gly Asp Tyr Val Ala Tyr
 305 310 315

cgc gta gga gag gtt gtc aac ggt gaa ggc gta agc tat cag 1001
 Arg Val Gly Glu Val Val Asn Gly Glu Gly Val Ser Tyr Gln
 320 325 330

tagtgaggat cc 1013

<210> 4

<211> 333

<212> PRT

<213> Arabidopsis thaliana

<400> 4

Met Asp Lys Asp Asp Asp Thr Asp Ser Leu Asn Tyr Lys Asp Ser Gly
 1 5 10 15

Val Asp Ile Asp Ala Gly Ala Glu Leu Val Lys Arg Ile Ala Lys Met

20				25				30							
Ala	Pro	Gly 35	Ile	Gly	Gly	Phe	Gly 40	Gly	Leu	Phe	Pro	Leu 45	Gly	Asp	Ser
Tyr	Leu 50	Val	Ala	Gly	Thr	Asp 55	Gly	Val	Gly	Thr	Lys 60	Leu	Lys	Leu	Ala
Phe 65	Glu	Thr	Gly	Ile	His 70	Asp	Thr	Ile	Gly	Ile 75	Asp	Leu	Val	Ala	Met 80
Ser	Val	Asn	Asp	Ile 85	Ile	Thr	Ser	Gly	Ala 90	Lys	Pro	Leu	Phe	Phe 95	Leu
Asp	Tyr	Phe	Ala 100	Thr	Ser	Arg	Leu	Asp 105	Val	Asp	Leu	Ala	Glu 110	Lys	Val
Ile	Lys	Gly 115	Ile	Val	Glu	Gly	Cys 120	Arg	Gln	Ser	Glu	Cys 125	Ala	Leu	Leu
Gly	Gly 130	Glu	Thr	Ala	Glu	Met 135	Pro	Asp	Phe	Tyr	Ala 140	Glu	Gly	Glu	Tyr
Asp 145	Leu	Ser	Gly	Phe	Ala 150	Val	Gly	Ile	Val	Lys 155	Lys	Thr	Ser	Val	Ile 160
Asn	Gly	Lys	Asn	Ile 165	Val	Ala	Gly	Asp	Val 170	Leu	Ile	Gly	Leu	Pro 175	Ser
Ser	Gly	Val	His 180	Ser	Asn	Gly	Phe	Ser 185	Leu	Val	Arg	Arg	Val 190	Leu	Ala
Arg	Ser	Asn 195	Leu	Ser	Leu	Asn	Asp 200	Ala	Leu	Pro	Gly	Gly 205	Ser	Ser	Thr
Leu	Gly 210	Asp	Ala	Leu	Met	Ala 215	Pro	Thr	Val	Ile	Tyr 220	Val	Lys	Gln	Val
Leu 225	Asp	Met	Ile	Glu	Lys 230	Gly	Gly	Val	Lys	Gly 235	Leu	Ala	His	Ile	Thr 240
Gly	Gly	Gly	Phe	Thr 245	Asp	Asn	Ile	Pro	Arg 250	Val	Phe	Pro	Asp	Gly 255	Leu
Gly	Ala	Val	Ile	His 260	Thr	Asp	Thr	Trp 265	Glu	Leu	Pro	Pro	Leu	Phe	Lys
Trp	Ile	Gln 275	Gln	Thr	Gly	Arg	Ile 280	Glu	Asp	Ser	Glu	Met 285	Arg	Arg	Thr
Phe	Asn 290	Leu	Gly	Ile	Gly	Met 295	Val	Met	Val	Val	Ser 300	Pro	Glu	Ala	Ala
Ser 305	Arg	Ile	Leu	Glu	Glu 310	Val	Lys	Asn	Gly	Asp 315	Tyr	Val	Ala	Tyr	Arg 320
Val	Gly	Glu	Val	Val 325	Asn	Gly	Glu	Gly	Val 330	Ser	Tyr	Gln			

5NSDOCID: <WO 9967402A2_1_>

- 7 -

<220>
<223> Description of Artificial Sequence:
oligonucleotide JG-L

<400> 5
gtacctcgag tctagactcg ag 22

<210> 6
<211> 28
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence:
oligonucleotide AS-1

<400> 6
gatcgagctc gttctcttct gtgtcatc 28

<210> 7
<211> 28
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence:
oligonucleotide AS-2

<400> 7
gatcccatgg tccccaggta aagacgtc 28

<210> 8
<211> 36
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence:
oligonucleotide slp242

<400> 8
cgcggtacct cactactgat agcttacgcc ttcacc 36

<210> 9
<211> 26
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence:
oligonucleotide slp244

<400> 9
ttgaagccat ggaagctcgg attttg 26

<210> 10
<211> 37
<212> DNA
<213> Artificial Sequence

- 8 -

<220>

<223> Description of Artificial Sequence:
oligonucleotide slp243

<400> 10

cgcatgccat ggataaagat gatgacactg atagtct

37



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12N 15/82, 15/54, 15/55, 9/10, 9/14, 5/10, C12Q 1/68, G01N 33/50, A01H 5/00		A2	(11) International Publication Number: WO 99/38986
			(43) International Publication Date: 5 August 1999 (05.08.99)
(21) International Application Number: PCT/EP99/00556		(74) Agent: BECKER, Konrad; Novartis AG, Corporate Intellectual Property, Patent & Trademark Dept., CH-4002 Basel (CH).	
(22) International Filing Date: 28 January 1999 (28.01.99)			
(30) Priority Data: 60/109,810 30 January 1998 (30.01.98) US		(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).	
(71) Applicant (for all designated States except AT US): NOVARTIS AG [CH/CH]; Schwarzwaldallee 215, CH-4058 Basel (CH).			
(71) Applicant (for AT only): NOVARTIS-ERFINDUNGEN VERWALTUNGSGESELLSCHAFT M.B.H. [AT/AT]; Brunner Strasse 59, A-1235 Vienna (AT).			
(72) Inventors; and (75) Inventors/Applicants (for US only): GUYER, Charles, David [CA/US]; 4615 J Hope Valley Road, Durham, NC 27707 (US). JOHNSON, Marie, Ann [US/US]; 408 Heather Drive, Raleigh, NC 27606 (US). VOLRATH, Sandra, Lynn [US/US]; 4225 Pine Oak Drive, Durham, NC 27707 (US). BRUNN, Sandra, Alice [CA/US]; 523 Sioux Lane, San Jose, CA 95123 (US). WARD, Eric, Russell [US/US]; 3761 Bentley Drive, Durham, NC 27707 (US).		Published Without international search report and to be republished upon receipt of that report.	
(54) Title: RIBOFLAVIN BIOSYNTHESIS GENES FROM PLANTS AND USES THEREOF			
(57) Abstract The present invention provides plant riboflavin biosynthesis genes, including a gene that encodes the β subunit of the plant riboflavin synthase enzyme complex (lumazine synthase) and a gene that encodes the bifunctional enzyme GTP cyclohydrolase II/DHBP synthase. Also disclosed are the recombinant production of these plant riboflavin biosynthesis enzymes in heterologous hosts, screening chemicals for herbicidal activity using these recombinantly produced enzymes, and the use of thereby identified herbicidal chemicals to suppress the growth of undesired vegetation. Furthermore, the present invention provides methods for the development of herbicide tolerance in plants, plant tissues, plant seeds and plant cells using the riboflavin biosynthesis genes of the invention.			

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Larvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

RIBOFLAVIN BIOSYNTHESIS GENES FROM PLANTS AND USES THEREOF

The invention relates generally to enzymatic activity involved in riboflavin biosynthesis in plants. In particular, the invention relates to plant genes that encode the bifunctional GTP cyclohydrolase II / DHBP synthase enzyme and the β subunit of the riboflavin synthase enzyme complex (lumazine synthase). The invention has various utilities, including the recombinant production of these riboflavin biosynthesis enzymes in heterologous hosts, the screening of chemicals for herbicidal activity, and the use of thereby identified herbicidal chemicals to control the growth of undesired vegetation. The invention may also be applied to the development of herbicide tolerance in plants, plant tissues, plant seeds, and plant cells.

I. Riboflavin Biosynthesis

Riboflavin (vitamin B₂ -- 6,7-dimethyl-9-(1-D-ribityl)-isoalloxazine) is synthesized by all plants and many microorganisms. Riboflavin is essential to basic metabolism because it is a precursor to coenzymes such as FAD and FMN, which are required in the enzymatic oxidation of carbohydrates. Biosynthesis of riboflavin starts from guanosine-5'-triphosphate (GTP) and proceeds through several enzymatic steps, as outlined in Figure 1 of Mironov *et al.*, *Mol. Gen. Genet.* 242:201-208 (1994), incorporated herein by reference.

GTP cyclohydrolase II is the first enzyme of riboflavin biosynthesis, catalyzing the synthesis of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate from GTP. DHBP synthase catalyzes the conversion of ribulose-5-phosphate to 3,4-dihydroxy-2-butanone phosphate (DHBP). In *Bacillus*, these two enzymatic activities are carried out by a single, bifunctional enzyme; in *E. coli*, however, these two enzymatic activities are carried out by two separate enzymes.

The riboflavin synthase protein is an approximately 1,000,000-Da enzyme complex consisting of approximately 60 β subunits and three α subunits. The β subunits form a capsid that catalyzes the conversion of 2,4-dioxy-5-amino-6-ribitylamino-pyrimidine (DARP) and 3,4-dihydroxy-2-butanone phosphate (DHBP) to 6,7-dimethyl-8-ribityllumazine (lumazine); hence, the β subunit is also known as "lumazine synthase". The α subunits, contained inside the β subunit capsid, then catalyze the conversion of two units of lumazine to one DARP molecule, which is recycled back into the first riboflavin synthase reaction, and one riboflavin molecule.

II. Herbicide Discovery

The use of herbicides to control undesirable vegetation such as weeds in crop fields has become almost a universal practice. The herbicide market exceeds 15 billion dollars annually. Despite this extensive use, weed control remains a significant and costly problem for farmers.

Effective use of herbicides requires sound management. For instance, the time and method of application and stage of weed plant development are critical to getting good weed control with herbicides. Since various weed species are resistant to herbicides, the production of effective new herbicides becomes increasingly important. Novel herbicides can now be discovered using high-throughput screens that implement recombinant DNA technology. Metabolic enzymes essential to plant growth and development can be recombinantly produced through standard molecular biological techniques and utilized as herbicide targets in screens for novel inhibitors of the enzymes' activity. The novel inhibitors discovered through such screens may then be used as herbicides to control undesirable vegetation.

III. Herbicide Tolerant Plants

Herbicides that exhibit greater potency, broader weed spectrum, and more rapid degradation in soil can also, unfortunately, have greater crop phytotoxicity. One solution applied to this problem has been to develop crops that are resistant or tolerant to herbicides. Crop hybrids or varieties tolerant to the herbicides allow for the use of the herbicides to kill weeds without attendant risk of damage to the crop. Development of tolerance can allow application of a herbicide to a crop where its use was previously precluded or limited (*e.g.* to pre-emergence use) due to sensitivity of the crop to the herbicide. For example, U.S. Patent No. 4,761,373 to Anderson *et al.* is directed to plants resistant to various imidazolinone or sulfonamide herbicides. The resistance is conferred by an altered acetohydroxyacid synthase (AHAS) enzyme. U.S. Patent No. 4,975,374 to Goodman *et al.* relates to plant cells and plants containing a gene encoding a mutant glutamine synthetase (GS) resistant to inhibition by herbicides that were known to inhibit GS, *e.g.* phosphinothricin and methionine sulfoximine. U.S. Patent No. 5,013,659 to Bedbrook *et al.* is directed to plants expressing a mutant acetolactate synthase that renders the plants resistant to inhibition by sulfonylurea herbicides. U.S. Patent No. 5,162,602 to Somers *et al.* discloses plants tolerant to inhibition by cyclohexanedione and

aryloxyphenoxypropanoic acid herbicides. The tolerance is conferred by an altered acetyl coenzyme A carboxylase (ACCase).

DEFINITIONS

For clarity, certain terms used in the specification are defined and presented as follows:

Activatable DNA Sequence: a DNA sequence that regulates the expression of genes in a genome, desirably the genome of a plant. The activatable DNA sequence is complementary to a target gene endogenous in the genome. When the activatable DNA sequence is introduced and expressed in a cell, it inhibits expression of the target gene. An activatable DNA sequence useful in conjunction with the present invention includes those encoding or acting as dominant inhibitors, such as a translatable or untranslatable sense sequence capable of disrupting gene function in stably transformed plants to positively identify one or more genes essential for normal growth and development of a plant. A preferred activatable DNA sequence is an antisense DNA sequence. The target gene preferably encodes a protein, such as a biosynthetic enzyme, receptor, signal transduction protein, structural gene product, or transport protein that is essential to the growth or survival of the plant. In an especially preferred embodiment, the target gene encodes lumazine synthase or the bifunctional enzyme GTP cyclohydrolase II / DHBP synthase. The interaction of the antisense sequence and the target gene results in substantial inhibition of the expression of the target gene so as to kill the plant, or at least inhibit normal plant growth or development.

Activatable DNA Construct: a recombinant DNA construct comprising a synthetic promoter operatively linked to the activatable DNA sequence, which when introduced into a cell, desirably a plant cell, is not expressed, i.e. is silent, unless a complete hybrid transcription factor capable of binding to and activating the synthetic promoter is present. The activatable DNA construct is introduced into cells, tissues, or plants to form stable transgenic lines capable of expressing the activatable DNA sequence.

Chimeric: "chimeric" is used to indicate that a DNA sequence, such as a vector or a gene, is comprised of more than one DNA sequences of distinct origin which are fused together by recombinant DNA techniques resulting in a DNA sequence, which does not occur naturally, and which particularly does not occur in the plant to be transformed.

DNA shuffling: DNA shuffling is a method to introduce mutations or rearrangements, preferably randomly, in a DNA molecule or to generate exchanges of DNA sequences between two or more DNA molecules, preferably randomly. The DNA molecule resulting from DNA shuffling is a shuffled DNA molecule that is a non-naturally occurring DNA molecule derived from at least one template DNA molecule. The shuffled DNA encodes an enzyme modified with respect to the enzyme encoded by the template DNA, and preferably has an altered biological activity with respect to the enzyme encoded by the template DNA.

Enzyme activity: means herein the ability of an enzyme to catalyze the conversion of a substrate into a product. A substrate for the enzyme comprises the natural substrate of the enzyme but also comprises analogues of the natural substrate which can also be converted by the enzyme into a product or into an analogue of a product. The activity of the enzyme is measured for example by determining the amount of product in the reaction after a certain period of time, or by determining the amount of substrate remaining in the reaction mixture after a certain period of time. The activity of the enzyme is also measured by determining the amount of an unused co-factor of the reaction remaining in the reaction mixture after a certain period of time or by determining the amount of used co-factor in the reaction mixture after a certain period of time. The activity of the enzyme is also measured by determining the amount of a donor of free energy or energy-rich molecule (e.g. ATP, phosphoenolpyruvate, acetyl phosphate or phosphocreatine) remaining in the reaction mixture after a certain period of time or by determining the amount of a used donor of free energy or energy-rich molecule (e.g. ADP, pyruvate, acetate or creatine) in the reaction mixture after a certain period of time.

Expression refers to the transcription and/or translation of an endogenous gene or a transgene in plants. In the case of antisense constructs, for example, expression may refer to the transcription of the antisense DNA only.

Gene refers to a coding sequence and associated regulatory sequences wherein the coding sequence is transcribed into RNA such as mRNA, rRNA, tRNA, snRNA, sense RNA or antisense RNA. Examples of regulatory sequences are promoter sequences, 5' and 3' untranslated sequences and

Herbicide: a chemical substance used to kill or suppress the growth of plants, plant cells, plant seeds, or plant tissues.

Heterologous DNA Sequence: a DNA sequence not naturally associated with a host cell into which it is introduced, including non-naturally occurring multiple copies of a naturally occurring DNA sequence.

Homologous DNA Sequence: a DNA sequence naturally associated with a host cell into which it is introduced.

Inhibitor: a chemical substance that inactivates the enzymatic activity of a protein such as a biosynthetic enzyme, receptor, signal transduction protein, structural gene product, or transport protein that is essential to the growth or survival of the plant. In the context of the instant invention, an inhibitor is a chemical substance that inactivates the enzymatic activity of lumazine synthase or the bifunctional enzyme GTP cyclohydrolase II / DHBP synthase from a plant. The term “herbicide” is used herein to define an inhibitor when applied to plants, plant cells, plant seeds, or plant tissues.

Isolated: in the context of the present invention, an isolated DNA molecule or an isolated enzyme is a DNA molecule or enzyme, by the hand of man, exists apart from its native environment and is therefore not a product of nature. An isolated DNA molecule or enzyme may exist in a purified form or may exist in a non-native environment such as, for example, a transgenic host cell.

Minimal Promoter: promoter elements, particularly a TATA element, that are inactive or that have greatly reduced promoter activity in the absence of upstream activation. In the presence of a suitable transcription factor, the minimal promoter functions to permit transcription.

Modified Enzyme Activity: enzyme activity different from that which naturally occurs in a plant (i.e. enzyme activity that occurs naturally in the absence of direct or indirect manipulation of such activity by man), which is tolerant to inhibitors that inhibit the naturally occurring enzyme activity.

Plant refers to any plant, particularly to seed plants

Plant cell: structural and physiological unit of the plant, comprising a protoplast and a cell wall. The plant cell may be in form of an isolated single cell or a cultured cell, or as a part of higher organized unit such as, for example, a plant tissue, or a plant organ.

Recombinant DNA: molecule a combination of DNA sequences that are joined together using recombinant DNA technology

Recombinant DNA technology: procedures used to join together DNA sequences as described, for example, in Sambrook et al., 1989, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press

Significant Increase: an increase in enzymatic activity that is larger than the margin of error inherent in the measurement technique, preferably an increase by about 2-fold or greater of the activity of the wild-type enzyme in the presence of the inhibitor, more preferably an increase by about 5-fold or greater, and most preferably an increase by about 10-fold or greater.

Significantly less: means that the amount of a product of an enzymatic reaction is larger than the margin of error inherent in the measurement technique, preferably a decrease by about 2-fold or greater of the activity of the wild-type enzyme in the absence of the inhibitor, more preferably an decrease by about 5-fold or greater, and most preferably an decrease by about 10-fold or greater.

Substantially Similar: in the context of the present invention, a DNA molecule that has at least 60 percent sequence identity with the portion of SEQ ID NO:1 that codes for lumazine synthase, i.e. that portion of SEQ ID NO:1 that encodes the amino acid sequence of SEQ ID NO:2; or a DNA molecule that has at least 60 percent sequence identity with the portion of SEQ ID NO:13 that codes for the bifunctional GTP cyclohydrolase II / DHBP synthase enzyme from a plant, i.e. that portion of SEQ ID NO:13 that encodes the amino acid sequence of SEQ ID NO:14. A substantially similar lumazine synthase nucleotide sequence hybridizes specifically to SEQ ID NO:1 or fragments thereof under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C. A substantially similar plant GTP cyclohydrolase II / DHBP synthase nucleotide sequence hybridizes specifically to SEQ ID NO:13 or fragments thereof under the above conditions. With respect to proteins, "substantially similar" as used herein means a protein sequence that is at least 90% identical to either the amino acid sequence set forth in SEQ ID NO:2 or the amino acid sequence set forth in SEQ ID NO:14.

Substrate: a substrate is the molecule that the enzyme naturally recognizes and converts to a product in the biochemical pathway in which the enzyme naturally carries out its function, or is a modified version of the molecule, which is also recognized by the enzyme and is converted by the enzyme to a product in an enzymatic reaction similar to the naturally-occurring reaction.

Synthetic refers to a nucleotide sequence comprising structural characters that are not present in the natural sequence. For example, an artificial sequence that resembles more closely the G+C content and the normal codon distribution of dicot and/or monocot genes is said to be synthetic.

Tolerance: the ability to continue normal growth or function when exposed to an inhibitor or herbicide.

Transformation: a process for introducing heterologous DNA into a cell, tissue, or plant. Transformed cells, tissues, or plants are understood to encompass not only the end product of a transformation process, but also transgenic progeny thereof.

Transgenic: stably transformed with a recombinant DNA molecule that preferably comprises a suitable promoter operatively linked to a DNA sequence of interest.

In view of the above, one object of the invention is to provide methods for identifying new or improved herbicides. Another object of the invention is to provide methods for using such new or improved herbicides to suppress the growth of plants such as weeds. Still another object of the invention is to provide improved crop plants that are tolerant to such new or improved herbicides.

In furtherance of these and other objects, the present invention provides a DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme involved in riboflavin biosynthesis, wherein the enzyme has either lumazine synthase activity or bifunctional GTP cyclohydrolase II / DHBP synthase activity.

According to one embodiment, the present invention provides a DNA molecule comprising a nucleotide sequence isolated from a plant that encodes the β subunit of riboflavin synthase (lumazine synthase). For example, the DNA molecule of the invention may comprises a nucleotide sequence that encodes an enzyme having lumazine synthase activity, wherein the enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:2. In another example, the DNA molecule of the invention comprises a nucleotide sequence that encodes an enzyme having lumazine synthase activity, wherein the enzyme comprises the amino acid sequence set forth in SEQ ID NO:2. In an other example, the DNA molecule of the invention comprises a nucleotide sequence isolated from a plant that encodes an enzyme having lumazine synthase activity, wherein said DNA molecule hybridizes to a DNA molecule that encodes the amino acid

sequence set forth in SEQ ID NO:2 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C. The invention further provides a DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme involved in riboflavin biosynthesis, wherein the enzyme has lumazine synthase activity, wherein said DNA molecule hybridizes to the coding sequence set forth in SEQ ID NO:1 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C. In yet another example, the DNA molecule of the invention comprises a nucleotide sequence that is substantially similar to the coding sequence set forth in SEQ ID NO:1 and that encodes an enzyme having lumazine synthase activity. In a further example, the DNA molecule of the invention comprises a nucleotide sequence isolated from a plant that encodes an enzyme having lumazine synthase activity, wherein said DNA molecule comprises a 20 base pair nucleotide portion identical in sequence to a consecutive 20 base pair portion of the coding sequence set forth in SEQ ID NO:1. In still another example, the DNA molecule of the invention comprises the coding sequence set forth in SEQ ID NO:1 and encodes an enzyme having lumazine synthase activity. Although the nucleotide sequence provided in SEQ ID NO:1 that encodes lumazine synthase was isolated from *Arabidopsis thaliana*, using the information provided by the present invention, the nucleotide sequence that encodes an enzyme having lumazine synthase activity can be obtained from any plant using standard methods known in the art.

According to another embodiment, the present invention provides a DNA molecule comprising a nucleotide sequence isolated from a plant that encodes the bifunctional GTP cyclohydrolase II / DHBP synthase. For example, the DNA molecule of the invention may comprise a nucleotide sequence that encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity, wherein the enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:14. In another example, the DNA molecule of the invention comprises a nucleotide sequence that encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity, wherein the enzyme comprises the amino acid sequence set forth in SEQ ID NO:14. In another example of the invention the DNA molecule comprises a nucleotide sequence isolated from a plant that encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity, wherein said DNA molecule hybridizes to a DNA molecule that

encodes the amino acid sequence set forth in SEQ ID NO: 14 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C. The invention further provides a DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme involved in riboflavin biosynthesis, wherein the enzyme has bifunctional GTP cyclohydrolase II / DHBP synthase activity, wherein said DNA molecule hybridizes to the coding sequence set forth in SEQ ID NO:13 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C.

In yet another example, the DNA molecule of the invention comprises a nucleotide sequence that is substantially similar to the coding sequence set forth in SEQ ID NO:13 and that encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity. In a further example, the DNA molecule of the invention comprises a nucleotide sequence isolated from a plant that encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity, wherein said DNA molecule comprises a 20 base pair nucleotide portion identical in sequence to a consecutive 20 base pair portion of the coding sequence set forth in SEQ ID NO:13. In still another example, the DNA molecule of the invention comprises the coding sequence set forth in SEQ ID NO:13 and encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity. Although the nucleotide sequence provided in SEQ ID NO:13 that encodes the bifunctional GTP cyclohydrolase II / DHBP synthase was isolated from *Arabidopsis thaliana*, using the information provided by the present invention, the nucleotide sequence that encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity can be obtained from any plant using standard methods known in the art.

The present invention also provides a chimeric gene comprising a promoter operatively linked to a DNA molecule of the invention. Further, the present invention provides a recombinant vector comprising such a chimeric gene, wherein the vector is capable of being stably transformed into a host cell. Still further, the present invention provides a host cell comprising such a vector, wherein the host cell is capable of expressing the DNA molecule encoding an enzyme involved in riboflavin biosynthesis. A host cell according to the invention may be a bacterial cell, a yeast cell, or a plant cell. Especially the host cell according to the invention may be a bacterial cell.

The present invention further provides a process for producing nucleotides sequences encoding gene products having altered lumazine synthase activity comprising: (a) shuffling a DNA molecule from a plant that encodes an enzyme having lumazine synthase activity, (b) expressing the resulting shuffled nucleotide sequences, and (c) selecting for altered lumazine synthase activity as compared to the activity of an enzyme encoded by the unshuffled DNA molecule. Preferably, the nucleotide sequence shuffled according to this method is SEQ ID NO: 1. The invention is also directed to a shuffled DNA molecule obtainable by this process. Preferably, the shuffled DNA molecule encodes an enzyme having enhanced tolerance to an inhibitor of lumazine synthase activity. The present invention also provides a chimeric gene comprising a promoter operatively linked to a shuffled DNA molecule; a recombinant vector comprising said chimeric gene, wherein said vector is capable of being stably transformed into a host cell; a host cell comprising said vector. Said host cell is preferably a bacterial cell, a yeast cell, or a plant cell, especially a plant cell. The invention is also directed to a plant or seed comprising such a plant cell. Preferably, said plant is tolerant to an inhibitor of lumazine synthase activity.

The present invention further provides a process for producing nucleotides sequences encoding gene products having altered bifunctional GTP cyclohydrolase II / DHBP synthase activity comprising: (a) shuffling a DNA molecule from a plant that encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity, (b) expressing the resulting shuffled nucleotide sequences, and (c) selecting for altered bifunctional GTP cyclohydrolase II / DHBP synthase activity as compared to the activity of an enzyme encoded by the unshuffled DNA molecule. Preferably, the nucleotide sequence shuffled according to this method is SEQ ID NO: 13. The invention is also directed to a shuffled DNA molecule obtainable by this process. Preferably, the shuffled DNA molecule encodes an enzyme having enhanced tolerance to an inhibitor of bifunctional GTP cyclohydrolase II / DHBP synthase activity. The present invention also provides a chimeric gene comprising a promoter operatively linked to a shuffled DNA molecule; a recombinant vector comprising said chimeric gene, wherein said vector is capable of being stably transformed into a host cell; a host cell comprising said vector. Said host cell is preferably a bacterial cell, a yeast cell, or a plant cell, especially a plant cell. The invention is also directed to a plant or seed comprising such a plant cell. Preferably, said plant is tolerant to an inhibitor of bifunctional GTP cyclohydrolase II / DHBP synthase activity.

In accordance with another embodiment, the present invention also relates to the recombinant production of the above-described riboflavin biosynthesis enzymes and

methods of use thereof. In particular, the present invention provides an isolated plant enzyme involved in riboflavin biosynthesis, wherein the enzyme has lumazine synthase activity. Preferably this enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:2. More preferably, this enzyme comprises the amino acid sequence set forth in SEQ ID NO:2. The present invention also provides an isolated plant enzyme involved in riboflavin biosynthesis, wherein the enzyme has bifunctional GTP cyclohydrolase II / DHBP synthase activity. Preferably, this enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:14. More preferably, this enzyme comprises the amino acid sequence set forth in SEQ ID NO:14.

The present invention further provides methods of using purified plant riboflavin biosynthesis enzymes such as lumazine synthase and bifunctional GTP cyclohydrolase II / DHBP synthase to screen for novel inhibitors thereof, which can then be used as herbicides to suppress the growth of undesirable vegetation in fields where crops are grown, particularly agronomically important crops such as maize and other cereal crops such as wheat, oats, rye, sorghum, rice, barley, millet, turf and forage grasses, and the like, as well as cotton, sugar cane, sugar beet, oilseed rape, and soybeans.

With regard to lumazine synthase, such a screen for chemicals having the ability to inhibit lumazine synthase activity preferably comprises the steps of: (a) combining an enzyme having lumazine synthase activity in a first reaction mixture with 2,4-dioxy-5-amino-6-ribitylamino-pyrimidine and 3,4-dihydroxy-2-butanone phosphate under conditions in which the enzyme is capable of catalyzing the synthesis of lumazine; (b) combining the chemical and the enzyme in a second reaction mixture with 2,4-dioxy-5-amino-6-ribitylamino-pyrimidine and 3,4-dihydroxy-2-butanone phosphate under the same conditions as in the first reaction mixture; (c) determining the amounts of lumazine produced in the first and second reaction mixtures; and (d) comparing the amounts of lumazine produced in the first and second reaction mixtures; wherein the chemical is capable of inhibiting the lumazine synthase activity of the enzyme if the amount of lumazine produced in the second reaction mixture is significantly less than the amount of lumazine produced in the first reaction mixture. Preferred is a method for screening according to the invention wherein the first reaction mixture comprises 50 μ M 2,4-dioxy-5-amino-6-ribitylamino-pyrimidine, and 0.5 mM 3,4-dihydroxy-2-butanone phosphate. Further preferred is a method for screening according to the invention, wherein the amounts of lumazine produced in the reaction mixtures are determined using a fluorimeter at an excitation wavelength of 407 nm.

With regard to the bifunctional GTP cyclohydrolase II / DHBP synthase, such a screen for chemicals having the ability to inhibit GTP cyclohydrolase II / DHBP synthase activity preferably comprises the steps of: (a) combining an enzyme having GTP cyclohydrolase II / DHBP synthase activity in a first reaction mixture with GTP or ribulose-5-phosphate under conditions in which the enzyme is capable of catalyzing the synthesis of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate, respectively; (b) combining the chemical and the enzyme in a second reaction mixture with GTP or ribulose-5-phosphate under the same conditions as in the first reaction mixture; (c) determining the amounts of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate produced in the first and second reaction mixtures; and (d) comparing the amounts of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate produced in the first and second reaction mixtures; wherein the chemical is capable of inhibiting the bifunctional GTP cyclohydrolase II / DHBP synthase activity of the enzyme if the amount of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate produced in the second reaction mixture is significantly less than the amount of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate produced in the first reaction mixture.

The present invention also embodies herbicidal chemicals identified by the above screening methods in addition to methods for suppressing the growth of plants by applying such herbicidal chemicals to the plants, whereby the chemicals inhibit the activity of lumazine synthase or bifunctional GTP cyclohydrolase II / DHBP synthase in the plants.

The present invention further embodies plants, plant tissues, plant seeds, and plant cells that have modified riboflavin biosynthesis enzyme activity and that are therefore tolerant to inhibition by a herbicide at levels normally inhibitory to naturally occurring riboflavin biosynthesis enzyme activity. Herbicide tolerant plants encompassed by the invention include those that would otherwise be potential targets for normally inhibiting herbicides, particularly the agronomically important crops mentioned above. According to this embodiment, plants, plant tissue, plant seeds, or plant cells are stably transformed with a recombinant DNA molecule comprising a suitable promoter functional in plants operatively linked to a nucleotide coding sequence that encodes a modified riboflavin biosynthesis enzyme that is tolerant to inhibition by a herbicide at a concentration that would normally inhibit the activity of wild-type, unmodified riboflavin biosynthesis enzyme. Modified

riboflavin biosynthesis enzyme activity may also be conferred upon a plant by increasing expression of wild-type herbicide-sensitive riboflavin biosynthesis enzyme by providing multiple copies of wild-type riboflavin biosynthesis genes to the plant or by overexpression of wild-type riboflavin biosynthesis genes under control of a stronger-than-wild-type promoter. The transgenic plants, plant tissue, plant seeds, or plant cells thus created are then selected by conventional selection techniques, whereby herbicide tolerant lines are isolated, characterized, and developed. Alternately, random or site-specific mutagenesis may be used to generate herbicide tolerant lines.

Therefore, the present invention provides a plant, plant cell, plant seed, or plant tissue comprising a DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme involved in riboflavin biosynthesis, wherein the enzyme has lumazine synthase activity and wherein the DNA molecule confers upon the plant, plant cell, plant seed, or plant tissue tolerance to a herbicide in amounts that normally naturally occurring lumazine synthase activity. According to one example of this embodiment, the enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:2. According to another example of this embodiment, the DNA molecule is substantially similar to the coding sequence set forth in SEQ ID NO:1. In a related aspect, the present invention is directed to a method for selectively suppressing the growth of weeds in a field containing a crop of planted crop seeds or plants, comprising the steps of: (a) planting herbicide tolerant crops or crop seeds, which are plants or plant seeds that are tolerant to a herbicide that inhibits naturally occurring lumazine synthase activity; and (b) applying to the crops or crop seeds and the weeds in the field a herbicide in amounts that inhibit naturally occurring lumazine synthase activity, wherein the herbicide suppresses the growth of the weeds without significantly suppressing the growth of the crops.

The present invention further provides a plant, plant cell, plant seed, or plant tissue comprising a DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme involved in riboflavin biosynthesis, wherein the enzyme has bifunctional GTP cyclohydrolase II / DHBP synthase activity and wherein the DNA molecule confers upon the plant, plant cell, plant seed, or plant tissue tolerance to a herbicide in amounts that normally naturally occurring bifunctional GTP cyclohydrolase II / DHBP synthase activity. According to one example of this embodiment, the enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:14. According to another example of this embodiment, the DNA molecule is

substantially similar to the coding sequence set forth in SEQ ID NO:13. In a related aspect, the present invention is directed to a method for selectively suppressing the growth of weeds in a field containing a crop of planted crop seeds or plants, comprising the steps of: (a) planting herbicide tolerant crops or crop seeds, which are plants or plant seeds that are tolerant to a herbicide that inhibits naturally occurring bifunctional GTP cyclohydrolase II / DHBP synthase activity; and (b) applying to the crops or crop seeds and the weeds in the field a herbicide in amounts that inhibit naturally occurring bifunctional GTP cyclohydrolase II / DHBP synthase activity, wherein the herbicide suppresses the growth of the weeds without significantly suppressing the growth of the crops.

Other objects and advantages of the present invention will become apparent to those skilled in the art from a study of the following description of the invention and non-limiting examples.

I. Plant Riboflavin Biosynthesis Genes

In one aspect, the present invention is directed to a DNA molecule comprising a nucleotide sequence isolated from a plant source that encodes the β subunit of riboflavin synthase (lumazine synthase). In particular, the present invention provides a DNA molecule isolated from *Arabidopsis thaliana* that encodes lumazine synthase and DNA molecules substantially similar thereto that encode enzymes having lumazine synthase activity. The DNA coding sequence for lumazine synthase from *Arabidopsis thaliana* is provided in SEQ ID NO:1.

In another aspect, the present invention is directed to a DNA molecule comprising a nucleotide sequence isolated from a plant source that encodes the bifunctional enzyme GTP cyclohydrolase II / 3,4-dihydroxy-2-butanone phosphate (DHBP). In particular, the present invention provides a DNA molecule isolated from *Arabidopsis thaliana* that encodes this bifunctional enzyme and DNA molecules substantially similar thereto that encode enzymes having GTP cyclohydrolase II / DHBP synthase activity. The DNA coding sequence for GTP cyclohydrolase II / DHBP synthase from *Arabidopsis thaliana* is provided in SEQ ID NO:13. The present invention represents the first recognition that in plants, GTP cyclohydrolase II and DHBP synthase constitute a single, bifunctional enzyme.

Based on Applicants' disclosure of the present invention, DNA sequences encoding riboflavin biosynthesis enzymes can, for the first time, be isolated from the genome of any desired plant species. An exemplary method for isolating riboflavin biosynthesis genes from

plants is described in Examples 1 and 11. With this method, searches of the *Arabidopsis thaliana* Expressed Sequence Tag (EST) database (Arabidopsis Biological Resource Center at Ohio State, Ohio State University, Columbus, OH) revealed ESTs with homologies to the *E. coli* riboflavin synthase β subunit and *B. subtilis* GTP cyclohydrolase. DNA fragments generated by PCR with primers specific to these ESTs were used to probe an *Arabidopsis* lambda ZAP library, whereupon cDNAs were isolated. The determined protein sequence encoded by one cDNA showed approximately 68% similarity to both the *E. coli* and *B. subtilis* riboflavin synthase β subunit. The determined protein sequence encoded by another cDNA showed approximately 70% similarity to the *B. subtilis* GTP cyclohydrolase.

Alternatively, riboflavin biosynthesis gene sequences can be isolated from any plant according to well known techniques based on their sequence similarity to the *Arabidopsis thaliana* coding sequences (SEQ ID NOs:1 and 13) taught by the present invention. In these techniques, all or part of a known plant riboflavin biosynthesis gene's coding sequence is used as a probe that selectively hybridizes to other riboflavin biosynthesis gene sequences present in a population of cloned genomic DNA fragments or cDNA fragments (i.e. genomic or cDNA libraries) from a chosen plant. Such techniques include hybridization screening of plated DNA libraries (either plaques or colonies; see, e.g., Sambrook *et al.*, "Molecular Cloning", eds., Cold Spring Harbor Laboratory Press. (1989)) and amplification by PCR using oligonucleotide primers corresponding to sequence domains conserved among known riboflavin biosynthesis enzyme's amino acid sequences (see, e.g. Innis *et al.*, "PCR Protocols, a Guide to Methods and Applications", pub. by Academic Press (1990)). These methods are particularly well suited to the isolation of riboflavin biosynthesis gene sequences from organisms closely related to the organism from which the probe sequence is derived. Thus, application of these methods using the *Arabidopsis* coding sequences as probes would be expected to be particularly well suited for the isolation of riboflavin biosynthesis gene sequences from other plant species, including monocotyledons and dicotyledons.

The isolated riboflavin biosynthesis gene sequences taught by the present invention can be manipulated according to standard genetic engineering techniques to suit any desired purpose. For example, an entire plant riboflavin biosynthesis gene sequence or portions thereof may be used as a probe capable of specifically hybridizing to coding sequences and messenger RNAs. To achieve specific hybridization under a variety of

conditions, such probes include sequences that are unique among plant riboflavin biosynthesis gene sequences and are at least 10 nucleotides in length, preferably at least 20 nucleotides in length, and most preferably at least 50 nucleotides in length. Such probes may be used to amplify and analyze riboflavin biosynthesis gene sequences from a chosen organism via PCR. This technique may be useful to isolate additional riboflavin biosynthesis gene sequences from a desired organism or as a diagnostic assay to determine the presence of riboflavin biosynthesis gene sequences in an organism. This technique may also be used to detect the presence of altered riboflavin biosynthesis gene sequences associated with a particular condition of interest such as herbicide tolerance, poor health, etc.

Lumazine synthase-specific and GTP cyclohydrolase II / DHBP synthase-specific hybridization probes can also be used to map the location of these native genes in the genome of a chosen plant using standard techniques based on the selective hybridization of the probe to genomic sequences. These techniques include, but are not limited to, identification of DNA polymorphisms identified or contained within the probe sequence, and use of such polymorphisms to follow segregation of the gene relative to other markers of known map position in a mapping population derived from self fertilization of a hybrid of two polymorphic parental lines (see e.g. Helentjaris *et al.*, *Plant Mol. Biol.* 5: 109 (1985); Sommer *et al.* *Biotechniques* 12:82 (1992); D'Ovidio *et al.*, *Plant Mol. Biol.* 15: 169 (1990)). While any plant riboflavin biosynthesis gene sequence is contemplated to be useful as a probe for mapping riboflavin biosynthesis genes, preferred probes are those gene sequences from plant species more closely related to the chosen plant species, and most preferred probes are those gene sequences from the chosen plant species. Mapping of riboflavin biosynthesis genes in this manner is contemplated to be particularly useful for breeding purposes. For instance, by knowing the genetic map position of a mutant riboflavin biosynthesis gene that confers herbicide resistance, flanking DNA markers can be identified from a reference genetic map (see, e.g., Helentjaris, *Trends Genet.* 3: 217 (1987)). During introgression of the herbicide resistance trait into a new breeding line, these markers can then be used to monitor the extent of linked flanking chromosomal DNA still present in the recurrent parent after each round of back-crossing.

Lumazine synthase-specific and GTP cyclohydrolase II / DHBP synthase-specific hybridization probes can also be used to quantify levels of riboflavin biosynthesis gene mRNA in a plant using standard techniques such as Northern blot analysis. This technique is useful as a diagnostic assay to detect altered levels of riboflavin biosynthesis gene

expression that are associated with particular conditions such as enhanced tolerance to herbicides that target riboflavin biosynthesis genes.

II. Essentiality of Riboflavin Biosynthesis Genes in Plants Demonstrated by Antisense Inhibition

As shown in the examples below, the essentiality of riboflavin biosynthesis genes for normal plant growth and development has been demonstrated by antisense inhibition of lumazine synthase in plants using the antisense validation system described in co-owned and co-pending application serial no. 08/978,650 [entitled "Methods and Compositions Useful for the Activation of Silent Transgenes", filed Nov. 26, 1997], incorporated herein by reference. In this system, a hybrid transcription factor gene is made that comprises a DNA-binding domain and an activation domain. In addition, an activatable DNA construct is made that comprises a synthetic promoter operatively linked to an activatable DNA sequence. The hybrid transcription factor gene and synthetic promoter are selected or designed such that the DNA binding domain of the hybrid transcription factor is capable of binding specifically to the synthetic promoter, which then activates expression of the activatable DNA sequence. A first plant is transformed with the hybrid transcription factor gene, and a second plant is transformed with the activatable DNA construct. The first plant and second plants are crossed to produce a progeny plant containing both the sequence encoding the hybrid transcription factor and the synthetic promoter, wherein the activatable DNA sequence is expressed in the progeny plant. In the preferred embodiment, the activatable DNA sequence is an antisense sequence capable of inactivating expression of an endogenous gene such as the lumazine synthase gene or the bifunctional GTP cyclohydrolase II / DHBP synthase gene. Hence, the progeny plant will be unable to normally express the endogenous gene.

This antisense validation system is especially useful for allowing expression of traits that might otherwise be unrecoverable as constitutively driven transgenes. For instance, foreign genes with potentially lethal effect or antisense genes or dominant-negative mutations designed to abolish function of essential genes, while of great interest in basic studies of plant biology, present inherent experimental problems. Decreased transformation frequencies are often cited as evidence of lethality associated with a particular constitutively driven transgene, but negative results of this type are laden with alternative trivial explanations. The present invention is an important advancement in the field of agriculture because it allows stable maintenance and propagation of a test transgene separate from its

expression. This ability to separate transgene insertion from expression is especially useful for firm conclusions about essentiality of gene function to be drawn. A substantial benefit of the present invention is that plant genes essential for normal growth or development can thus be identified in this manner. The identification of such genes provide useful targets for screening compound libraries for effective herbicides. Below, the antisense validation system is described in greater detail:

A. Hybrid Transcription Factor Gene

A hybrid transcription factor gene for use in the antisense validation system described herein comprises DNA sequences encoding (1) a DNA-binding domain and (2) an activation domain that interacts with components of transcriptional machinery assembling at a promoter. Gene fragments are joined, typically such that the DNA binding domain is toward the 5' terminus and the activator domain is toward the 3' terminus, to form a hybrid gene whose expression produces a hybrid transcription factor. One skilled in the art is capable of routinely combining various DNA sequences encoding DNA binding domains with various DNA sequences encoding activation domains to produce a wide array of hybrid transcription factor genes. Examples of DNA sequences encoding DNA binding domains include, but are not limited to, those encoding the DNA binding domains of GAL4, bacteriophage 434, *lexA*, *lacI*, and phage lambda repressor. Examples of DNA sequences encoding the activation domain include, but are not limited to, those encoding the acidic activation domains of herpes simplex VP16, maize C1, and P1. In addition, suitable activation domains can be isolated by fusing DNA pieces from an organism of choice to a suitable DNA binding domain and selecting directly for function (Estruch *et al.*, (1994) *Nucleic Acids Res.* 22: 3983-3989). Domains of transcriptional activator proteins can be swapped between proteins of diverse origin (Brent and Ptashne (1985) *Cell* 43: 729-736). A preferable hybrid transcription factor gene comprises DNA sequences encoding the GAL4 DNA binding domain fused to the maize C1 activation domain.

B. Activatable DNA Construct

An activatable DNA construct for use in the antisense validation system described herein comprises (1) a synthetic promoter operatively linked to (2) an activatable DNA sequence. The synthetic promoter comprises at least one DNA binding site recognized by the DNA binding domain of the hybrid transcription factor, and a minimal promoter, preferably a TATA element derived from a promoter recognized by plant cells. More

particularly the TATA element is derived from a promoter recognized by the plant cell type into which the synthetic promoter will be incorporated. Desirably, the DNA binding site is repeated multiple times in the synthetic promoter so that the minimal promoter may be more effectively activated, such that the activatable DNA sequence associated with the synthetic promoter is more effectively expressed. One skilled in the art can use routine molecular biology and recombinant DNA technology to make desirable synthetic promoters. Examples of DNA binding sites that can be used to make synthetic promoters useful in the invention include, but are not limited to, the upstream activating sequence (UAS_G) recognized by the GAL4 DNA binding protein, the *lac* operator, and the *lexA* binding site. Examples of promoter TATA elements recognized by plant cells include those derived from CaMV 35S, the maize *Bz1* promoter, and the UBQ3 promoter. An especially preferable synthetic promoter comprises a truncated CaMV 35S sequence containing the TATA element (nucleotides -59 to +48 relative to the start of transcription), fused at its 5' end to approximately 10 concatemeric direct repeats of the upstream activating sequence (UAS_G) recognized by the GAL4 DNA binding domain.

The activatable DNA sequence encompasses any DNA sequence for which stable introduction and expression in a plant cell is desired. Particularly desirable activatable DNA sequences are sense or antisense sequences, whose expression results in decreased expression of their endogenous counterpart genes, thereby inhibiting normal plant growth or development. The activatable DNA sequence is operatively linked to the synthetic promoter to form the activatable DNA construct. The activatable DNA sequence in the activatable DNA construct is not expressed, i.e. is silent, in transgenic lines, unless a hybrid transcription factor capable of binding to and activating the synthetic promoter, is also present. The activatable DNA construct subsequently is introduced into cells, tissues or plants to form stable transgenic lines expressing the activatable DNA sequence, as described more fully below. In the context of the present invention, the activatable DNA sequence preferably comprises an antisense lumazine synthase sequence or an antisense bifunctional GTP cyclohydrolase II / DHBP synthase sequence.

C. Transgenic Plants Containing the Hybrid Transcription Factor Gene or the Activatable DNA Construct

The antisense validation system described herein utilizes a first plant containing the hybrid transcription factor gene and a second plant containing the activatable DNA construct. The hybrid transcription factor genes and activatable DNA constructs described

above are introduced into the plants by methods well known and routinely used in the art, including but not limited to crossing, *Agrobacterium*-mediated transformation, Ti plasmid vectors, direct DNA uptake such as microprojectile bombardment, liposome mediated uptake, micro-injection, etc. Transformants are screened for the presence and functionality of the transgenes according to standard methods known to those skilled in the art.

D. Transgenic Plants Containing Both the Hybrid Transcription Factor Gene and the Activatable DNA Construct

F1 plants containing both the hybrid transcription factor gene and the activatable DNA construct are generated by cross-pollination and selected for the presence of an appropriate marker. In contrast to plants containing the activatable DNA construct alone, the F1 plants generate high levels of activatable DNA sequence expression product, comparable to those obtained with strong constitutive promoters such as CaMV 35S.

Antisense Validation Assay:

Thus, a useful assay in the system described herein comprises the following steps:

- a) providing a first transgenic plant stably transformed with a hybrid transcription factor gene encoding a hybrid transcription factor capable of activating a synthetic promoter when said synthetic promoter is present in the plant, wherein the first transgenic plant is homozygous for the hybrid transcription factor;
- b) providing a second transgenic plant stably transformed with an activatable DNA construct comprising a synthetic promoter activatable by the hybrid transcription factor of step a) operatively linked to an activatable DNA sequence, such as an antisense lumazine synthase sequence or an antisense GTP cyclohydrolase II / DHBP synthase sequence;
- c) crossing the first transgenic plant with the second transgenic plant to yield F1 plants expressing the activatable DNA sequence in the presence of the hybrid transcription factor; and
- d) determining the effect of expression of the activatable DNA sequence on the F1 plants.

III. Recombinant Production of Plant Riboflavin Biosynthesis Enzymes and Uses Thereof

For recombinant production of a plant riboflavin biosynthesis enzyme in a host organism, a plant riboflavin biosynthesis coding sequence of the invention may be inserted into an expression cassette designed for the chosen host and introduced into the host where it is recombinantly produced. The choice of specific regulatory sequences such as

promoter, signal sequence, 5' and 3' untranslated sequences, and enhancer appropriate for the chosen host is within the level of skill of the routineer in the art. The resultant molecule, containing the individual elements linked in proper reading frame, may be inserted into a vector capable of being transformed into the host cell. Suitable expression vectors and methods for recombinant production of proteins are well known for host organisms such as *E. coli*, yeast, and insect cells (see, e.g., Luckow and Summers, *Bio/Technol.* 6: 47 (1988)). Specific examples include plasmids such as pBluescript (Stratagene, La Jolla, CA), pFLAG (International Biotechnologies, Inc., New Haven, CT), pTrcHis (Invitrogen, La Jolla, CA), and baculovirus expression vectors, e.g., those derived from the genome of *Autographica californica* nuclear polyhedrosis virus (AcMNPV). A preferred baculovirus/insect system is pVI11392/Sf21 cells (Invitrogen, La Jolla, CA).

Recombinantly produced plant riboflavin biosynthesis enzymes can be isolated and purified using a variety of standard techniques. The actual techniques that may be used will vary depending upon the host organism used, whether the enzyme is designed for secretion, and other such factors familiar to the skilled artisan (see, e.g. chapter 16 of Ausubel, F. *et al.*, "Current Protocols in Molecular Biology", pub. by John Wiley & Sons, Inc. (1994)).

Recombinantly produced plant riboflavin biosynthesis enzymes are useful for a variety of purposes. For example, they can be used in *in vitro* assays to screen known herbicidal chemicals whose target has not been identified to determine if they inhibit riboflavin biosynthesis enzymes. Such *in vitro* assays may also be used as more general screens to identify chemicals that inhibit such enzymatic activity and that are therefore herbicide candidates. Alternatively, recombinantly produced riboflavin biosynthesis enzymes may be used to further characterize their association with known inhibitors in order to rationally design new inhibitory herbicides as well as herbicide tolerant forms of the enzymes.

Inhibitor Assay:

Thus, an assay useful for identifying inhibitors of essential plant genes, such as plant riboflavin biosynthesis genes, comprises the steps of:

- a) reacting a plant riboflavin biosynthesis enzyme and a substrate thereof in the presence of a suspected inhibitor of the enzyme's function;

- b) comparing the rate of enzymatic activity in the presence of the suspected inhibitor to the rate of enzymatic activity under the same conditions in the absence of the suspected inhibitor; and
- c) determining whether the suspected inhibitor inhibits the riboflavin biosynthesis enzyme.

For example, the inhibitory effect on plant lumazine synthase may be determined by a reduction or complete inhibition of lumazine synthesis in the assay. Such a determination may be made by comparing, in the presence and absence of the candidate inhibitor, the amount of lumazine synthesized in the *in vitro* assay using fluorescence or absorbance detection as described *infra* in the Examples. A similar assay may be used to screen for inhibitors of the bifunctional plant GTP cyclohydrolase II / DHBP synthase enzyme.

In addition, recombinantly produced plant riboflavin biosynthesis enzymes may be used to elucidate the complex structure of these molecules, such as has been done for riboflavin synthase from *Bacillus subtilis* (Ladenstein, *et al.*, (1988) *J. Mol. Biol.* 203, 1045-1070). Such information regarding the structure of the plant riboflavin biosynthesis enzymes may be used, for example, in the rational design of new inhibitory herbicides.

IV. Herbicide Tolerant Plants

The present invention is further directed to plants, plant tissue, plant seeds, and plant cells tolerant to herbicides that inhibit the naturally occurring riboflavin biosynthesis in these plants, wherein the tolerance is conferred by altered riboflavin biosynthesis enzyme activity. Altered riboflavin biosynthesis enzyme activity may be conferred upon a plant according to the invention by increasing expression of wild-type herbicide-sensitive riboflavin biosynthesis enzyme by providing additional wild-type riboflavin biosynthesis genes to the plant, by expressing modified herbicide-tolerant riboflavin biosynthesis enzymes in the plant, or by a combination of these techniques. Representative plants include any plants to which these herbicides are applied for their normally intended purpose. Preferred are agronomically important crops such as cotton, soybean, oilseed rape, sugar beet, maize, rice, wheat, barley, oats, rye, sorghum, millet, turf, forage, turf grasses, and the like.

A. Increased Expression of Wild-Type Riboflavin Biosynthesis Enzymes

Achieving altered riboflavin biosynthesis enzyme activity through increased expression results in a level of a riboflavin biosynthesis enzyme in the plant cell at least sufficient to overcome growth inhibition caused by the herbicide. The level of expressed

enzyme generally is at least two times, preferably at least five times, and more preferably at least ten times the natively expressed amount. Increased expression may be due to multiple copies of a wild-type riboflavin biosynthesis gene; multiple occurrences of the coding sequence within the gene (*i.e.* gene amplification) or a mutation in the non-coding, regulatory sequence of the endogenous gene in the plant cell. Plants having such altered gene activity can be obtained by direct selection in plants by methods known in the art (see, *e.g.* U.S. Patent No. 5,162,602, and U.S. Patent No. 4,761,373, and references cited therein). These plants also may be obtained by genetic engineering techniques known in the art. Increased expression of a herbicide-sensitive riboflavin biosynthesis gene can also be accomplished by stably transforming a plant cell with a recombinant or chimeric DNA molecule comprising a promoter capable of driving expression of an associated structural gene in a plant cell operatively linked to a homologous or heterologous structural gene encoding the riboflavin biosynthesis enzyme.

B. Expression of Modified Herbicide-Tolerant Riboflavin Biosynthesis Enzymes

According to this embodiment, plants, plant tissue, plant seeds, or plant cells are stably transformed with a recombinant DNA molecule comprising a suitable promoter functional in plants operatively linked to a coding sequence encoding a herbicide tolerant form of a riboflavin biosynthesis enzyme. A herbicide tolerant form of the enzyme has at least one amino acid substitution, addition or deletion that confers tolerance to a herbicide that inhibits the unmodified, naturally occurring form of the enzyme. The transgenic plants, plant tissue, plant seeds, or plant cells thus created are then selected by conventional selection techniques, whereby herbicide tolerant lines are isolated, characterized, and developed. Below are described methods for obtaining genes that encode herbicide tolerant forms of riboflavin biosynthesis enzymes:

One general strategy involves direct or indirect mutagenesis procedures on microbes. For instance, a genetically manipulatable microbe such as *E. coli* or *S. cerevisiae* may be subjected to random mutagenesis *in vivo* with mutagens such as UV light or ethyl or methyl methane sulfonate. Mutagenesis procedures are described, for example, in Miller, *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1972); Davis *et al.*, *Advanced Bacterial Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1980); Sherman *et al.*, *Methods in Yeast Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1983); and U.S. Patent No. 4,975,374. The microbe

selected for mutagenesis contains a normal, inhibitor-sensitive riboflavin biosynthesis gene and is dependent upon the activity conferred by this gene. The mutagenized cells are grown in the presence of the inhibitor at concentrations that inhibit the unmodified gene. Colonies of the mutagenized microbe that grow better than the unmutagenized microbe in the presence of the inhibitor (i.e. exhibit resistance to the inhibitor) are selected for further analysis. Riboflavin biosynthesis genes from these colonies are isolated, either by cloning or by PCR amplification, and their sequences are elucidated. Sequences encoding altered gene products are then cloned back into the microbe to confirm their ability to confer inhibitor tolerance.

A method of obtaining mutant herbicide-tolerant alleles of a plant riboflavin biosynthesis gene involves direct selection in plants. For example, the effect of a mutagenized riboflavin biosynthesis gene on the growth inhibition of plants such as *Arabidopsis*, soybean, or maize is determined by plating seeds sterilized by art-recognized methods on plates on a simple minimal salts medium containing increasing concentrations of the inhibitor. Such concentrations are in the range of 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 110, 300, 1000 and 3000 parts per million (ppm). The lowest dose at which significant growth inhibition can be reproducibly detected is used for subsequent experiments.

Mutagenesis of plant material is utilized to increase the frequency at which resistant alleles occur in the selected population. Mutagenized seed material is derived from a variety of sources, including chemical or physical mutagenesis or seeds, or chemical or physical mutagenesis or pollen (Neuffer, In *Maize for Biological Research* Sheridan, ed. Univ. Press, Grand Forks, ND., pp. 61-64 (1982)), which is then used to fertilize plants and the resulting M₁ mutant seeds collected. Typically for *Arabidopsis*, M₂ seeds (Lehle Seeds, Tucson, AZ), which are progeny seeds of plants grown from seeds mutagenized with chemicals, such as ethyl methane sulfonate, or with physical agents, such as gamma rays or fast neutrons, are plated at densities of up to 10,000 seeds/plate (10 cm diameter) on minimal salts medium containing an appropriate concentration of inhibitor to select for tolerance. Seedlings that continue to grow and remain green 7-21 days after plating are transplanted to soil and grown to maturity and seed set. Progeny of these seeds are tested for tolerance to the herbicide. If the tolerance trait is dominant, plants whose seed segregate 3:1 / resistant:sensitive are presumed to have been heterozygous for the resistance at the M₂ generation. Plants that give rise to all resistant seed are presumed to have been homozygous for the resistance at the M₂ generation. Such mutagenesis on

intact seeds and screening of their M2 progeny seed can also be carried out on other species, for instance soybean (see, e.g. U.S. Pat. No. 5,084,082). Alternatively, mutant seeds to be screened for herbicide tolerance are obtained as a result of fertilization with pollen mutagenized by chemical or physical means.

Confirmation that the genetic basis of the herbicide tolerance is a modified riboflavin biosynthesis gene is ascertained as exemplified below. First, alleles of the riboflavin biosynthesis gene from plants exhibiting resistance to the inhibitor are isolated using PCR with primers based either upon conserved regions in the *Arabidopsis* cDNA coding sequences shown in SEQ ID NO:1 or SEQ ID NO:13 or, more preferably, based upon the unaltered riboflavin biosynthesis gene sequence from the plant used to generate tolerant alleles. After sequencing the alleles to determine the presence of mutations in the coding sequence, the alleles are tested for their ability to confer tolerance to the inhibitor on plants into which the putative tolerance-conferring alleles have been transformed. These plants can be either *Arabidopsis* plants or any other plant whose growth is susceptible to the inhibitors. Second, the riboflavin biosynthesis genes are mapped relative to known restriction fragment length polymorphisms (RFLPs) (See, for example, Chang *et al. Proc. Natl. Acad. Sci. USA* 85: 6856-6860 (1988); Nam *et al., Plant Cell* 1: 699-705 (1989). The tolerance trait is independently mapped using the same markers. When tolerance is due to a mutation in that riboflavin biosynthesis gene, the tolerance trait maps to a position indistinguishable from the position of the riboflavin biosynthesis gene.

Another method of obtaining herbicide-tolerant alleles of a riboflavin biosynthesis gene is by selection in plant cell cultures. Explants of plant tissue, e.g. embryos, leaf disks, etc. or actively growing callus or suspension cultures of a plant of interest are grown on medium in the presence of increasing concentrations of the inhibitory herbicide or an analogous inhibitor suitable for use in a laboratory environment. Varying degrees of growth are recorded in different cultures. In certain cultures, fast-growing variant colonies arise that continue to grow even in the presence of normally inhibitory concentrations of inhibitor. The frequency with which such faster-growing variants occur can be increased by treatment with a chemical or physical mutagen before exposing the tissues or cells to the inhibitor. Putative tolerance-conferring alleles of the riboflavin biosynthesis gene are isolated and tested as described in the foregoing paragraphs. Those alleles identified as conferring herbicide tolerance may then be engineered for optimal expression and transformed into

the plant. Alternatively, plants can be regenerated from the tissue or cell cultures containing these alleles.

Still another method involves mutagenesis of wild-type, herbicide sensitive plant riboflavin biosynthesis genes in bacteria or yeast, followed by culturing the microbe on medium that contains inhibitory concentrations of the inhibitor and then selecting those colonies that grow in the presence of the inhibitor. More specifically, a plant cDNA, such as the *Arabidopsis* cDNA encoding lumazine synthase (SEQ ID NO:1) or the bifunctional GTP cyclohydrolase II / DHBP synthase enzyme (SEQ ID NO:13) is cloned into a microbe that otherwise lacks the selected gene's activity. The transformed microbe is then subjected to *in vivo* mutagenesis or to *in vitro* mutagenesis by any of several chemical or enzymatic methods known in the art, e.g. sodium bisulfite (Shortle *et al.*, *Methods Enzymol.* 100:457-468 (1983); methoxylamine (Kadonaga *et al.*, *Nucleic Acids Res.* 13:1733-1745 (1985); oligonucleotide-directed saturation mutagenesis (Hutchinson *et al.*, *Proc. Natl. Acad. Sci. USA*, 83:710-714 (1986); or various polymerase misincorporation strategies (see, e.g. Shortle *et al.*, *Proc. Natl. Acad. Sci. USA*, 79:1588-1592 (1982); Shiraishi *et al.*, *Gene* 64:313-319 (1988); and Leung *et al.*, *Technique* 1:11-15 (1989). Colonies that grow in the presence of normally inhibitory concentrations of inhibitor are picked and purified by repeated restreaking. Their plasmids are purified and tested for the ability to confer tolerance to the inhibitor by retransforming them into the microbe lacking riboflavin biosynthesis gene activity. The DNA sequences of cDNA inserts from plasmids that pass this test are then determined.

Herbicide resistant riboflavin biosynthesis genes are also obtained using methods involving *in vitro* recombination, also called DNA shuffling. By DNA shuffling, mutations, preferably random mutations, are introduced in riboflavin biosynthesis genes. DNA shuffling also leads to the recombination and rearrangement of sequences within a riboflavin biosynthesis gene or to recombination and exchange of sequences between two or more different riboflavin biosynthesis protein encoding sequences. These methods allow for the production of millions of mutated riboflavin biosynthesis genes. The mutated genes, or shuffled genes, are screened for desirable properties, e.g. improved tolerance to herbicides and for mutations that provide broad spectrum tolerance to the different classes of inhibitor chemistry. Such screens are well within the skills of a routineer in the art.

In a preferred embodiment, a mutagenized riboflavin biosynthesis gene is formed from at least one template riboflavin biosynthesis gene, wherein the template riboflavin

biosynthesis gene has been cleaved into double-stranded random fragments of a desired size, and comprising the steps of adding to the resultant population of double-stranded random fragments one or more single or double-stranded oligonucleotides, wherein said oligonucleotides comprise an area of identity and an area of heterology to the double-stranded random fragments; denaturing the resultant mixture of double-stranded random fragments and oligonucleotides into single-stranded fragments; incubating the resultant population of single-stranded fragments with a polymerase under conditions which result in the annealing of said single-stranded fragments at said areas of identity to form pairs of annealed fragments, said areas of identity being sufficient for one member of a pair to prime replication of the other, thereby forming a mutagenized double-stranded polynucleotide; and repeating the second and third steps for at least two further cycles, wherein the resultant mixture in the second step of a further cycle includes the mutagenized double-stranded polynucleotide from the third step of the previous cycle, and the further cycle forms a further mutagenized double-stranded polynucleotide, wherein the mutagenized polynucleotide is a mutated riboflavin biosynthesis gene having enhanced tolerance to a herbicide which inhibits naturally occurring riboflavin biosynthesis activity. In a preferred embodiment, the concentration of a single species of double-stranded random fragment in the population of double-stranded random fragments is less than 1% by weight of the total DNA. In a further preferred embodiment, the template double-stranded polynucleotide comprises at least about 100 species of polynucleotides. In another preferred embodiment, the size of the double-stranded random fragments is from about 5 bp to 5 kb. In a further preferred embodiment, the fourth step of the method comprises repeating the second and the third steps for at least 10 cycles. Such method is described e.g. in Stemmer et al. (1994) Nature 370: 389-391, in US Patent 5,605,793 and in Crameri et al. (1998) Nature 391: 288-291, as well as in WO 97/20078, and these references are incorporated herein by reference.

In another preferred embodiment, any combination of two or more different riboflavin biosynthesis genes are mutagenized *in vitro* by a staggered extension process (StEP), as described e.g. in Zhao et al. (1998) Nature Biotechnology 16: 258-261. Briefly, the two or more riboflavin biosynthesis genes are used as template for PCR amplification with the extension cycles of the PCR reaction preferably carried out at a lower temperature than the optimal polymerization temperature of the polymerase. For example, when a thermostable polymerase with an optimal temperature of approximately 72°C is used, the temperature for the extension reaction is desirably below 72°C, more desirably below 65°C, preferably

below 60°C, more preferably the temperature for the extension reaction is 55°C. Additionally, the duration of the extension reaction of the PCR cycles is desirably shorter than usually carried out in the art, more desirably it is less than 30 seconds, preferably it is less than 15 seconds, more preferably the duration of the extension reaction is 5 seconds. Only a short DNA fragment is polymerized in each extension reaction, allowing template switch of the extension products between the starting DNA molecules after each cycle of denaturation and annealing, thereby generating diversity among the extension products. The optimal number of cycles in the PCR reaction depends on the length of the riboflavin biosynthesis coding regions to be mutagenized but desirably over 40 cycles, more desirably over 60 cycles, preferably over 80 cycles are used. Optimal extension conditions and the optimal number of PCR cycles for every combination of riboflavin biosynthesis genes are determined as described in using procedures well-known in the art. The other parameters for the PCR reaction are essentially the same as commonly used in the art. The primers for the amplification reaction are preferably designed to anneal to DNA sequences located outside of the coding sequence of the riboflavin biosynthesis genes, e.g. to DNA sequences of a vector comprising the riboflavin biosynthesis genes, whereby the different riboflavin biosynthesis genes used in the PCR reaction are preferably comprised in separate vectors. The primers desirably anneal to sequences located less than 500 bp away from riboflavin biosynthesis coding sequences, preferably less than 200 bp away from the riboflavin biosynthesis coding sequences, more preferably less than 120 bp away from the riboflavin biosynthesis coding sequences. Preferably, the riboflavin biosynthesis coding sequences are surrounded by restriction sites, which are included in the DNA sequence amplified during the PCR reaction, thereby facilitating the cloning of the amplified products into a suitable vector.

In another preferred embodiment, fragments of riboflavin biosynthesis genes having cohesive ends are produced as described in WO 98/05765. The cohesive ends are produced by ligating a first oligonucleotide corresponding to a part of a riboflavin biosynthesis gene to a second oligonucleotide not present in the gene or corresponding to a part of the gene not adjoining to the part of the gene corresponding to the first oligonucleotide, wherein the second oligonucleotide contains at least one ribonucleotide. A double-stranded DNA is produced using the first oligonucleotide as template and the second oligonucleotide as primer. The ribonucleotide is cleaved and removed. The nucleotide(s) located 5' to the ribonucleotide is also removed, resulting in double-stranded

fragments having cohesive ends. Such fragments are randomly reassembled by ligation to obtain novel combinations of gene sequences.

Any riboflavin biosynthesis gene or any combination of riboflavin biosynthesis genes is used for *in vitro* recombination in the context of the present invention, for example, a riboflavin biosynthesis gene derived from a plant, such as, e.g. *Arabidopsis thaliana*, e.g. a riboflavin biosynthesis gene set forth in SEQ ID NO:1 or SEQ ID NO:13, or a riboflavin biosynthesis gene from *Bacillus* or *E. coli*. Whole riboflavin biosynthesis genes or portions thereof are used in the context of the present invention. The library of mutated riboflavin biosynthesis genes obtained by the methods described above are cloned into appropriate expression vectors and the resulting vectors are transformed into an appropriate host, for example an algae like *Chlamydomonas*, a yeast or a bacteria. A preferred host is preferably a host that otherwise lacks riboflavin biosynthesis gene activity. Host cells transformed with the vectors comprising the library of mutated riboflavin biosynthesis genes are cultured on medium that contains inhibitory concentrations of the inhibitor and those colonies that grow in the presence of the inhibitor are selected. Colonies that grow in the presence of normally inhibitory concentrations of inhibitor are picked and purified by repeated restreaking. Their plasmids are purified and the DNA sequences of cDNA inserts from plasmids that pass this test are then determined.

An assay for identifying a modified riboflavin biosynthesis gene that is tolerant to an inhibitor may be performed in the same manner as the assay to identify inhibitors of the riboflavin biosynthesis enzyme (Inhibitor Assay, above) with the following modifications: First, a mutant riboflavin biosynthesis enzyme is substituted in one of the reaction mixtures for the wild-type riboflavin biosynthesis enzyme of the inhibitor assay. Second, an inhibitor of wild-type enzyme is present in both reaction mixtures. Third, mutated activity (activity in the presence of inhibitor and mutated enzyme) and unmutated activity (activity in the presence of inhibitor and wild-type enzyme) are compared to determine whether a significant increase in enzymatic activity is observed in the mutated activity when compared to the unmutated activity. Mutated activity is any measure of activity of the mutated enzyme while in the presence of a suitable substrate and the inhibitor. Unmutated activity is any measure of activity of the wild-type enzyme while in the presence of a suitable substrate and the inhibitor. A significant increase is defined as an increase in enzymatic activity that is larger than the margin of error inherent in the measurement technique, preferably an increase by about 2-fold or greater of the activity of the wild-type enzyme in the presence of

the inhibitor, more preferably an increase by about 5-fold or greater, most preferably an increase by about 10-fold or greater.

In addition to being used to create herbicide-tolerant plants, genes encoding herbicide tolerant riboflavin biosynthesis enzymes can also be used as selectable markers in plant cell transformation methods. For example, plants, plant tissue, plant seeds, or plant cells transformed with a transgene can also be transformed with a gene encoding an altered riboflavin biosynthesis enzyme capable of being expressed by the plant. The transformed cells are transferred to medium containing an inhibitor of the enzyme in an amount sufficient to inhibit the survivability of plant cells not expressing the modified gene wherein only the transformed cells will survive. The method is applicable to any plant cell capable of being transformed with a modified riboflavin biosynthesis enzyme-encoding gene, and can be used with any transgene of interest. Expression of the transgene and the modified gene can be driven by the same promoter functional in plant cells, or by separate promoters.

V. Plant Transformation Technology

A wild-type or herbicide-tolerant form of the riboflavin biosynthesis gene can be incorporated in plant or bacterial cells using conventional recombinant DNA technology. Generally, this involves inserting a DNA molecule encoding the riboflavin biosynthesis enzyme into an expression system to which the DNA molecule is heterologous (i.e., not normally present) using standard cloning procedures known in the art. The vector contains the necessary elements for the transcription and translation of the inserted protein-coding sequences in a host cell containing the vector. A large number of vector systems known in the art can be used, such as plasmids, bacteriophage viruses and other modified viruses. The components of the expression system may also be modified to increase expression. For example, truncated sequences, nucleotide substitutions or other modifications may be employed. Expression systems known in the art can be used to transform virtually any crop plant cell under suitable conditions. Transformed cells can be regenerated into whole plants such that the chosen form of the riboflavin biosynthesis gene confers herbicide tolerance in the transgenic plants.

A. Requirements for Construction of Plant Expression Cassettes

Gene sequences intended for expression in transgenic plants are first assembled in expression cassettes behind a suitable promoter expressible in plants. The expression

cassettes may also comprise any further sequences required or selected for the expression of the transgene. Such sequences include, but are not restricted to, transcription terminators, extraneous sequences to enhance expression such as introns, vital sequences, and sequences intended for the targeting of the gene product to specific organelles and cell compartments. These expression cassettes can then be easily transferred to the plant transformation vectors described *infra*. The following is a description of various components of typical expression cassettes.

1. Promoters

The selection of the promoter used in expression cassettes will determine the spatial and temporal expression pattern of the transgene in the transgenic plant. Selected promoters will express transgenes in specific cell types (such as leaf epidermal cells, mesophyll cells, root cortex cells) or in specific tissues or organs (roots, leaves or flowers, for example) and the selection will reflect the desired location of accumulation of the gene product. Alternatively, the selected promoter may drive expression of the gene under various inducing conditions. Promoters vary in their strength, i.e., ability to promote transcription. Depending upon the host cell system utilized, any one of a number of suitable promoters known in the art can be used. For example, for constitutive expression, the CaMV 35S promoter, the rice actin promoter, or the ubiquitin promoter may be used. For regulatable expression, the chemically inducible PR-1 promoter from tobacco or *Arabidopsis* may be used (*see, e.g.*, U.S. Patent No. 5,689,044).

2. Transcriptional Terminators

A variety of transcriptional terminators are available for use in expression cassettes. These are responsible for the termination of transcription beyond the transgene and its correct polyadenylation. Appropriate transcriptional terminators are those that are known to function in plants and include the CaMV 35S terminator, the *tml* terminator, the nopaline synthase terminator and the pea *rbcS* E9 terminator. These can be used in both monocotyledons and dicotyledons.

3. Sequences for the Enhancement or Regulation of Expression

Numerous sequences have been found to enhance gene expression from within the transcriptional unit and these sequences can be used in conjunction with the genes of this

invention to increase their expression in transgenic plants. For example, various intron sequences such as introns of the maize *Adhl* gene have been shown to enhance expression, particularly in monocotyledonous cells. In addition, a number of non-translated leader sequences derived from viruses are also known to enhance expression, and these are particularly effective in dicotyledonous cells.

4. Coding Sequence Optimization

The coding sequence of the selected gene may be genetically engineered by altering the coding sequence for optimal expression in the crop species of interest. Methods for modifying coding sequences to achieve optimal expression in a particular crop species are well known (see, *e.g.* Perlak *et al.*, *Proc. Natl. Acad. Sci. USA* 88: 3324 (1991); and Koziel *et al.*, *Bio/technol.* 11: 194 (1993)).

5. Targeting of the Gene Product Within the Cell

Various mechanisms for targeting gene products are known to exist in plants and the sequences controlling the functioning of these mechanisms have been characterized in some detail. For example, the targeting of gene products to the chloroplast is controlled by a signal sequence found at the amino terminal end of various proteins which is cleaved during chloroplast import to yield the mature protein (*e.g.* Comai *et al.* *J. Biol. Chem.* 263: 15104-15109 (1988)). Other gene products are localized to other organelles such as the mitochondrion and the peroxisome (*e.g.* Unger *et al.* *Plant Molec. Biol.* 13: 411-418 (1989)). The cDNAs encoding these products can also be manipulated to effect the targeting of heterologous gene products to these organelles. In addition, sequences have been characterized which cause the targeting of gene products to other cell compartments. Amino terminal sequences are responsible for targeting to the ER, the apoplast, and extracellular secretion from aleurone cells (Koehler & Ho, *Plant Cell* 2: 769-783 (1990)). Additionally, amino terminal sequences in conjunction with carboxy terminal sequences are responsible for vacuolar targeting of gene products (Shinshi *et al.* *Plant Molec. Biol.* 14: 357-368 (1990)). By the fusion of the appropriate targeting sequences described above to transgene sequences of interest it is possible to direct the transgene product to any organelle or cell compartment.

B. Construction of Plant Transformation Vectors

Numerous transformation vectors available for plant transformation are known to those of ordinary skill in the plant transformation arts, and the genes pertinent to this invention can be used in conjunction with any such vectors. The selection of vector will depend upon the preferred transformation technique and the target species for transformation. For certain target species, different antibiotic or herbicide selection markers may be preferred. Selection markers used routinely in transformation include the *nptII* gene, which confers resistance to kanamycin and related antibiotics (Messing & Vierra. Gene 19: 259-268 (1982); Bevan et al., Nature 304:184-187 (1983)), the *bar* gene, which confers resistance to the herbicide phosphinothricin (White et al., Nucl. Acids Res 18: 1062 (1990), Spencer et al. Theor. Appl. Genet 79: 625-631 (1990)), the *hph* gene, which confers resistance to the antibiotic hygromycin (Blochinger & Diggelmann, Mol Cell Biol 4: 2929-2931), and the *dhfr* gene, which confers resistance to methatrexate (Bourouis et al., EMBO J. 2(7): 1099-1104 (1983)), and the EPSPS gene, which confers resistance to glyphosate (U.S. Patent Nos. 4,940,935 and 5,188,642).

1. Vectors Suitable for *Agrobacterium* Transformation

Many vectors are available for transformation using *Agrobacterium tumefaciens*. These typically carry at least one T-DNA border sequence and include vectors such as pBIN19 (Bevan, Nucl. Acids Res. (1984)) and pXYZ. Typical vectors suitable for *Agrobacterium* transformation include the binary vectors pCIB200 and pCIB2001, as well as the binary vector pCIB10 and hygromycin selection derivatives thereof. (See, for example, U.S. Patent No. 5,639,949).

2. Vectors Suitable for non-*Agrobacterium* Transformation

Transformation without the use of *Agrobacterium tumefaciens* circumvents the requirement for T-DNA sequences in the chosen transformation vector and consequently vectors lacking these sequences can be utilized in addition to vectors such as the ones described above which contain T-DNA sequences. Transformation techniques that do not rely on *Agrobacterium* include transformation via particle bombardment, protoplast uptake (e.g. PEG and electroporation) and microinjection. The choice of vector depends largely on the preferred selection for the species being transformed. Typical vectors suitable for non-*Agrobacterium* transformation include pCIB3064, pSOG19, and pSOG35. (See, for example, U.S. Patent No. 5,639,949).

C. Transformation Techniques

Once the coding sequence of interest has been cloned into an expression system, it is transformed into a plant cell. Methods for transformation and regeneration of plants are well known in the art. For example, Ti plasmid vectors have been utilized for the delivery of foreign DNA, as well as direct DNA uptake, liposomes, electroporation, micro-injection, and microprojectiles. In addition, bacteria from the genus *Agrobacterium* can be utilized to transform plant cells.

Transformation techniques for dicotyledons are well known in the art and include *Agrobacterium*-based techniques and techniques that do not require *Agrobacterium*. Non-*Agrobacterium* techniques involve the uptake of exogenous genetic material directly by protoplasts or cells. This can be accomplished by PEG or electroporation mediated uptake, particle bombardment-mediated delivery, or microinjection. In each case the transformed cells are regenerated to whole plants using standard techniques known in the art.

Transformation of most monocotyledon species has now also become routine. Preferred techniques include direct gene transfer into protoplasts using PEG or electroporation techniques, particle bombardment into callus tissue, as well as *Agrobacterium*-mediated transformation.

VI. Breeding

The wild-type or altered form of a riboflavin biosynthesis gene of the present invention can be utilized to confer herbicide tolerance to a wide variety of plant cells, including those of gymnosperms, monocots, and dicots. Although the gene can be inserted into any plant cell falling within these broad classes, it is particularly useful in crop plant cells, such as rice, wheat, barley, rye, corn, potato, carrot, sweet potato, sugar beet, bean, pea, chicory, lettuce, cabbage, cauliflower, broccoli, turnip, radish, spinach, asparagus, onion, garlic, eggplant, pepper, celery, carrot, squash, pumpkin, zucchini, cucumber, apple, pear, quince, melon, plum, cherry, peach, nectarine, apricot, strawberry, grape, raspberry, blackberry, pineapple, avocado, papaya, mango, banana, soybean, tobacco, tomato, sorghum and sugarcane.

The high-level expression of a wild-type riboflavin biosynthesis gene and/or the expression of herbicide-tolerant forms of a riboflavin biosynthesis gene conferring herbicide tolerance in plants, in combination with other characteristics important for production and

quality, can be incorporated into plant lines through breeding approaches and techniques known in the art.

Where a herbicide tolerant riboflavin biosynthesis gene allele is obtained by direct selection in a crop plant or plant cell culture from which a crop plant can be regenerated, it is moved into commercial varieties using traditional breeding techniques to develop a herbicide tolerant crop without the need for genetically engineering the allele and transforming it into the plant.

The invention will be further described by reference to the following detailed examples. These examples are provided for purposes of illustration only, and are not intended to be limiting unless otherwise specified.

BRIEF DESCRIPTION OF THE SEQUENCES IN THE SEQUENCE LISTING

SEQ ID NO:1 is a cDNA sequence encoding the β subunit of riboflavin synthase (lumazine synthase) from *Arabidopsis thaliana*.

SEQ ID NO:2 is the predicted amino acid sequence of *Arabidopsis thaliana* lumazine synthase encoded by SEQ ID NO:1.

SEQ ID NO:3 is oligonucleotide DG-63.

SEQ ID NO:4 is oligonucleotide DG-65.

SEQ ID NO:5 is oligonucleotide JG-L.

SEQ ID NO:6 is oligonucleotide RS-1.

SEQ ID NO:7 is oligonucleotide RS-2.

SEQ ID NO:8 is a synthetic peptide used in Example 7.

SEQ ID NO:9 is a another synthetic peptide used in Example 7.

SEQ ID NO:10 is oligonucleotide DG-252.

SEQ ID NO:11 is oligonucleotide DG-253.

SEQ ID NO:12 is oligonucleotide DG-254.

SEQ ID NO:13 is a partial cDNA sequence encoding the bifunctional GTP cyclohydrolase II / DHBP synthase enzyme from *Arabidopsis thaliana*.

SEQ ID NO:14 is the predicted amino acid sequence of the mature *Arabidopsis thaliana* GTP cyclohydrolase II / DHBP synthase enzyme encoded by SEQ ID NO:13.

SEQ ID NO:15 is oligonucleotide DG-67.

SEQ ID NO:16 is oligonucleotide DG-69.

SEQ ID NO:17 is oligonucleotide DG-392a.

SEQ ID NO:18 is oligonucleotide DG-393a.

SEQ ID NO:19 is oligonucleotide DG-390a.

SEQ ID NO:20 is oligonucleotide DG-391a.

EXAMPLES

Standard recombinant DNA and molecular cloning techniques used here are well known in the art and are described by Sambrook, *et al.*, Molecular Cloning, eds., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (1989) and by T.J. Silhavy, M.L. Berman, and L.W. Enquist, Experiments with Gene Fusions, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1984) and by Ausubel, F.M. *et al.*, Current Protocols in Molecular Biology, pub. by Greene Publishing Assoc. and Wiley-Interscience (1987).

Example 1: Isolation of a cDNA Encoding Lumazine Synthase from *Arabidopsis*

A search of the *Arabidopsis thaliana* Expressed Sequence Tag (EST) database (Arabidopsis Biological Resource Center at Ohio State, Ohio State University, Columbus, OH) revealed an EST (EST # P25540, gb acc. # Z34233) with homology to the β Subunit of Riboflavin Synthase from *E.coli*. Using plasmid DNA of an *Arabidopsis* cDNA library (Minet *et al.*, (1992) *Plant J.* 2: 417-422) as a template, and synthetic oligonucleotides DG-63 (SEQ ID NO:3) and DG-65 (SEQ ID NO:4) designed to the EST sequence, a 204-bp DNA fragment was generated using the polymerase chain reaction (PCR). The 204-bp fragment was ligated into the TA cloning vector pCR II (Invitrogen Corp., San Diego, CA). Sequence determination by the chain termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA) confirmed that the sequence of the 204-bp fragment was identical to the sequence of EST #P25540.

Approximately 150,000 pfu of a lambda ZAP *Arabidopsis* cDNA library was plated at a density of 8,000 plaques per 10 cm Petri dish, and filter lifts of the plaques were made after 7 hours growth at 37°C. The plaque lifts were probed with the 204-bp fragment labeled with 32P-dCTP by the random priming method by means of a PrimeTime kit (International Biotechnologies, Inc., New Haven, CT). Hybridization conditions were 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA, 1% bovine albumin at 65°C. After hybridization overnight, the filters were washed with 1% SDS, 50mM NaPO₄, 1mM EDTA at 65°C. Six positively hybridizing plaques were detected by autoradiography. After purification to single plaques, cDNA inserts were isolated, and their sequences were determined by the chain termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA). A database search of the

longest clone, designated RS β -1, using the GAP program (Deveraux, *et al.*, (1984) *Nucleic Acids Res.* 12:387-95) revealed sequence similarity to the riboflavin synthase β subunit from *E. coli*. The proteins are 68% similar and 44% identical. In addition, a comparison of the *Arabidopsis* mature protein to the *E. coli* riboflavin synthase β subunit suggests a chloroplast transit peptide is present.

RS β -1, in the pBluescript SK vector, was deposited as pDG-4a.t. with the Agricultural Research Culture Collection (NRRL), 1815 N. University St., Peoria, IL 61604, USA under the terms of the Budapest Treaty on February 7, 1995, and assigned NRRL accession number B-21400.

The *Arabidopsis* cDNA sequence encoding RS β -1 is set forth in SEQ ID NO:1 and the encoded amino acid sequence is set forth in SEQ ID NO:2.

Example 2: Isolation of Additional Lumazine Synthase Genes based on Sequence Similarity to the *Arabidopsis* Lumazine Synthase Coding Sequence

A phage or plasmid library is plated at a density of approximately 8,000 pfu per 10 cm Petri dish, and filter lifts of the plaques are made after 7 hours growth at 37°C. The plaque lifts are probed with the cDNA set forth in SEQ ID NO:1, labeled with 32P-dCTP by the random priming method by means of a PrimeTime kit (International Biotechnologies, Inc., New Haven, CT). Hybridization conditions are 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C. After hybridization overnight, the filters are washed with 2X SSC, 1% SDS at 50°C. Positively hybridizing plaques are detected by autoradiography. After purification to single plaques, cDNA inserts are isolated, and their sequences determined by the chain termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA). This experimental protocol can be used by one of ordinary skill in the art to obtain lumazine synthase genes substantially similar to the *Arabidopsis* coding sequence (SEQ ID NO:1) from any other plant species.

Example 3: Construction of a Vector Containing a GAL4 Binding Site/Minimal 35S CaMV Promoter Fused to Antisense Lumazine Synthase

pAT71:

10 GAL4 binding sites and the minimal 35S promoter (-59 to +1) were excised from pGALLuc2 (Goff, *et al.*, (1991) *Genes & Development* 5: 298-309) as an *EcoRI-PstI* fragment and inserted into the respective sites of pLdescript, yielding pAT52. pAT66 was constructed with a three-way ligation between the *HindIII-PstI* fragment of pAT52, a *PstI-EcoRI* fragment of pCIB1716 (contains a 35S untranslated leader, GUS gene, 35S terminator) and *HindIII-EcoRI* cut pUC18. The 35S leader of pAT66 was excised with *PstI-NcoI* and replaced with a PCR-generated 35S leader extending from +1 to +48 to yield pAT71.

pJG304:

Plasmid pBS SK+ (Stratagene, LaJolla, CA) was linearized with *SacI*, treated with mung bean nuclease to remove the *SacI* site, and re-ligated with T4 ligase to make pJG201. The 10XGAL4 consensus binding site/CaMV 35S minimal promoter/GUS gene/CaMV terminator cassette was removed from pAT71 with *KpnI* and cloned into the *KpnI* site of pJG201 to make pJG304.

pJG304 was partially digested with restriction endonuclease *Asp718* to isolate a full-length linear fragment. This fragment was ligated with a molar excess of the 22 base oligonucleotide JG-L (SEQ ID NO:5). Restriction analysis was used to identify a clone with this linker inserted 5' to the GAL4 DNA binding site, and this plasmid was designated pJG304?XhoI.

pDG1:

A fragment of the lumazine synthase cDNA clone was PCR-amplified from the cDNA clone RS β -1 using the oligonucleotides RS-1 (SEQ ID NO:6) and RS-2 (SEQ ID NO:7). This PCR product comprises the 5' portion of the lumazine synthase cDNA (SEQ ID NO:1), ending at base pair 792.

The vector pJG304?XhoI was digested with *SacI* and *NcoI* to excise the GUS gene coding sequence. The lumazine synthase PCR fragment was digested with *SacI* and *NcoI* and ligated into pJG304?XhoI to make pDG1.

Example 4: Plant Transformation Vectors For Lumazine Synthase Antisense Expression
From The GAL4 Binding Site/CaMV Minimal 35S Promoter

pJG261:

Vector p⁺PTV (Becker, *et al.*, (1992) *Plant Molecular Biology* 20: 1195-1197) was digested with *EcoRI* and *HindIII* to remove the nopaline synthase promoter/GUS cassette. Concurrently, the superlinker was excised from pSE380 (Invitrogen, San Diego, CA) with *EcoRI* and *HindIII* and cloned into the *EcoRI*/*HindIII* linearized pGPTV, to make pJG261.

pDG2:

pDG1 was cut with *XhoI* to excise the cassette containing the GAL4 DNA binding site/35S minimal promoter/antisense lumazine synthase/CaMV terminator fusion. This cassette was ligated into *XhoI*-digested pJG261, such that transcription was divergent from that of the *bar* selectable marker, producing pDG2.

Example 5: Production Of GAL4 Binding Site/Minimal CaMV 35S
Antisense Lumazine Synthase Transgenic Plants

pDG2 was electro-transformed (Bio-Rad Laboratories, Hercules, CA) into *Agrobacterium tumefaciens* strain C58C1 (pMP90), and *Arabidopsis* plants (Ecotype Columbia) were transformed by infiltration (Bechtold, *et al.*, (1993) *C. R. Acad. Sci. Paris*, 316: 1188-93). Seeds from the infiltrated plants were selected on germination medium (Murashige-Skoog salts at 4.3 g/liter, Mes at 0.5 g/liter, 1% sucrose, thiamine at 10 ug/liter, pyridoxine at 5 ug/liter, nicotinic acid at 5 ug/liter, myo-inositol at 1 mg/liter, pH 5.8) containing Basta at 15 mg/liter.

Example 6: Production of GAL4/C1 Transactivator Transgenic Plants

pSGZL1 was constructed by ligating the GAL4-C1 *EcoRI* fragment from pGALC1 (Goff, *et al.*, (1991) *Genes & Development*, 5: 298-309) into the *EcoRI* site of pIC20H. The

GAL4-C1 fragment of pSGZL1 was excised with *Bam*HI-*Bgl*II and inserted into the *Bam*HI site of pCIB770 (Rothstein, *et al.*, (1987) *Gene* 53: 153-161) yielding pAT53.

Arabidopsis root explants were transformed with pAT53 as described in Valvekens, *et al.*, (1985) *PNAS USA* 85: 5536-5540. Transgenic plants with single site insertion and positive for GAL4/C1 expression were taken to homozygosity.

Example 7: Antisense Inhibition of Lumazine Synthase Using a GAL4/C1 Transactivator and a GAL4 Binding Site/Minimal CaMV 35S Promoter

Fifteen transgenic plants containing the GAL4 binding site/minimal CaMV 35S promoter/antisense lumazine synthase construct were transplanted to soil and grown to maturity in the greenhouse. Flowers borne on the primary transformants were crossed to pollen from the homozygous GAL4/C1 transactivator line pAT53-103. F1 seeds were plated on germination medium and germination medium containing 15 mg/liter Basta. One line gave a 50% lethal phenotype on plates. Seedlings from the remaining F1 lines were transplanted to soil and grown to maturity in the greenhouse. Half of the seedlings from 2 F1 lines died while in soil.

Lumazine synthase antibody was generated in goat by injecting the synthetic peptides CIGAVIRGDTT (SEQ ID NO:8) and KAGNKGAEALTALTEM (SEQ ID NO:9) conjugated to purified protein derivative. Western analysis of F1 plants revealed a significant decrease in lumazine synthase levels (Towbin *et al.*, *PNAS USA* 76: 4350-4354).

Example 8: Expression and Purification of Recombinant Plant Lumazine Synthase in *E. coli*

To produce recombinant plant lumazine synthase in *E. coli*, a translational fusion of the *Arabidopsis* lumazine synthase cDNA (SEQ ID NO:1) to the 5' end of the thioredoxin gene (LaVallie *et al.*, (1992) *Biotechnology* 11:187-193) was created in pET-32a (Novagen, Inc., Madison, WI) using PCR. Synthetic oligonucleotide primers DG-252 (SEQ ID NO:10), DG-253 (SEQ ID NO:11), and DG-254 (SEQ ID NO:12) were used in a polymerase chain reaction to amplify DNA fragments of 693-bp and 483-bp in length. The PCR products were digested with *Nco*I and *Eco*RI. The digestion products were separated on a low-gelling-temperature agarose gel and the fragments were excised. In parallel, plasmid pET32a was digested with *Nco*I and *Eco*RI. The digestion products were separated on a gel, and the

pET32a vector was excised from the gel. The vector fragment was ligated to the two PCR-generated fragments, and the ligation products were transformed into competent *E. coli* XL1 Blue cells (Stratagene, La Jolla, CA).

Ampicillin-resistant colonies were selected, cultured, and their plasmid DNAs extracted. The structures of the plasmids were confirmed by sequencing with the chain termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA). The recombinant plasmids with expected structure were designated pET32aRS β FL-1 and pET32aRS β No CTP-1.

Plasmids pET32aRS β FL-1 and pET32aRS β No CTP-1 were transformed into competent *E. coli* BL21(DE3) cells, and recombinant protein was expressed and purified according to the manufacturer's instructions (pET System Manual, Novagen, Inc., Madison, WI). The resulting fusion proteins produced by this strain contained approximately 132 amino acids of *E. coli* thioredoxin protein, His-Tag, and thrombin cleavage site, followed by the presumptive mature coding sequence for *Arabidopsis* lumazine synthase, which begins at codon 1 of the predicted protein coding sequence for plasmid pET32aRS β FL-1, and codon 71 of the predicted protein coding sequence for plasmid pET32aRS β No CTP-1.

Example 9: Lumazine Synthase Activity Assay

Lumazine synthase activity is detected using an HPLC and fluorimeter combination. Both lumazine and 2,4-dioxy-5-amino-6-ribitylamino-pyrimidine (DARP) are fluorescent under the following conditions: excitation wavelength 407 nm; emission wavelength 487 nm. However, lumazine is about 6-fold more fluorescent than an equimolar concentration of DARP. There is also a 6-fold difference in absorbance between lumazine and DARP at 405 nm. 3,4-dihydroxy-2-butanone phosphate does not fluoresce. Lumazine and DARP can be separated on a C18 column using 33% 90 mM formic acid, 60% water, and 7% methanol. Lumazine elutes first at four minutes, followed two minutes later by DARP.

The peak area can be directly related to the molar quantity of lumazine produced. Optimization studies have shown the buffer for the reaction to be preferably 100 mM KPO₄, pH 7, 5 mM β -mercaptoethanol, 2 mM DTT. The enzyme is active at a pH range of 6.5 - 7.5, but pH 7 is most preferable. Kinetic studies show that the K_m for the butanone phosphate is 190 μ M and the K_m for DARP is 5.5 μ M. Kis *et al.*, *Biochem.* 34: 2883-2892 (1995) reported K_m values of 130 and 5, respectively for the bacterial enzyme. The reaction

is incubated at 37°C for ten minutes and then stopped by the addition of 5% TCA. The precipitated proteins are removed by centrifugation and 10 µl of the supernatant is injected onto the HPLC. Because the reaction can proceed non-enzymatically, controls should be run with all samples to subtract this background activity.

Example 10: High Throughput Screen

A high throughput screen for novel inhibitors of lumazine synthase preferably exploits the fact that lumazine and DARP fluoresce at different intensities under optimal conditions for lumazine or the fact that there is a 6-fold difference in absorbance between these two compounds. An example of a protocol for a high throughput screen using fluorescence detection is as follows: lumazine synthase, buffer, test substance, and DARP are mixed together in the wells of a 96-well microtiter plate to a volume of 190 µl, and the initial fluorescence value is determined (with, for example, a Waters fluorimetric microtiter plate reader). Reactions commence with the addition of a 10 µl aliquot of 3,4-dihydroxy-2-butanone phosphate. After an appropriate incubation time, fluorescence is determined again. The differences between initial and final readings are then scaled as a percent of control reactions. Initial concentrations of substrates in the complete reaction mixture are preferably 50 µM for DARP and 0.5 mM for the butanone phosphate. Lumazine synthase amount and incubation time are adjusted to allow for the production of lumazine to a concentration of approximately 25 µM. This will produce a fluorescence signal that is approximately 3 to 4-fold greater than background.

Example 11: Isolation of a cDNA Encoding the Bifunctional GTP Cyclohydrolase II / 3,4-Dihydroxy-2-Butanone-4-Phosphate Synthase from *Arabidopsis*

A search of the *Arabidopsis thaliana* Expressed Sequence Tag (EST) database (Arabidopsis Biological Resource Center at Ohio State, Ohio State University, Columbus, OH) revealed an EST (EST # SCH1T7P; gb acc. # T12970) with homology to GTP cyclohydrolase from *Bacillus subtilis*. Using plasmid DNA of an *Arabidopsis* cDNA library (Minet et al, (1992) *Plant J.* 2. 417-422) as a template, and synthetic oligonucleotides DG-67 (SEQ ID NO:15) and DG-69 (SEQ ID NO:16) designed to the EST sequence, a 322-bp DNA fragment was generated using the polymerase chain reaction (PCR). The 322-bp

fragment was ligated into the TA cloning vector pCR II (Invitrogen Corp., San Diego, CA). Sequence determination by the chain termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA), confirmed that the sequence of the 322-bp fragment was identical to the sequence of EST # SCH1T7P.

Approximately 150,000 pfu of a lambda ZAP *Arabidopsis* cDNA library was plated at a density of 8,000 plaques per 10 cm Petri dish, and filter lifts of the plaques were made after 7 hours growth at 37°C. The plaque lifts were probed with the 322-bp fragment labeled with 32P-dCTP by the random priming method by means of a PrimeTime kit (International Biotechnologies, Inc., New Haven, CT). Hybridization conditions were 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA, 1% bovine albumin at 65°C. After hybridization overnight, the filters were washed with 1% SDS, 50mM NaPO₄, 1mM EDTA at 65°C. Ten positively hybridizing plaques were detected by autoradiography. After purification to single plaques, cDNA inserts were isolated, and their sequences were determined by the chain termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA). A database search of the longest clone, designated GTP-1, using the GAP program (Deveraux et al., Nucleic Acids Res. 12:387-95 (1984), revealed sequence similarity to the bifunctional GTP cyclohydrolase II/3,4-dihydroxy-2-butanone-4-phosphate synthase of *Bacillus subtilis*. The proteins are 70% similar and 54% identical. In addition, a comparison of the *Arabidopsis* mature protein to the *Bacillus subtilis* GTP cyclohydrolase II/3,4-dihydroxy-2-butanone-4-phosphate synthase suggests a chloroplast transit peptide is present.

GTP-1, in the pBluescript SK vector, was deposited as pDG-3a.t. with the Agricultural Research Culture Collection (NRRL), 1815 N. University St., Peoria, IL 61604, USA under the terms of the Budapest Treaty on February 7, 1995, and assigned NRRL accession number B-21399.

The *Arabidopsis* cDNA sequence encoding GTP-1 is set forth in SEQ ID NO:13 and the amino acid sequence of the encoded mature protein, without the putative transit peptide, is set forth in SEQ ID NO:14.

Example 12: Isolation of Additional GTP Cyclohydrolase II / 3,4-Dihydroxy-2-Butanone-4-Phosphate Synthase Genes Based On Sequence Homology the *Arabidopsis* GTP Cyclohydrolase II / 3,4-Dihydroxy-2-Butanone-4-Phosphate Synthase Coding Sequence

A phage or plasmid library is plated at a density of approximately 8,000 pfu per 10 cm Petri dish, and filter lifts of the plaques are made after 7 hours growth at 37°C. The plaque lifts are probed with the cDNA set forth in SEQ ID NO:13, labeled with 32P-dCTP by the random priming method by means of a PrimeTime kit (International Biotechnologies, Inc., New Haven, CT). Hybridization conditions are 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C. After hybridization overnight, the filters are washed with 2X SSC, 1% SDS at 50°C. Positively hybridizing plaques are detected by autoradiography. After purification to single plaques, cDNA inserts are isolated, and their sequences are determined by the chain termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA). This experimental protocol can be used by one of ordinary skill in the art to obtain bifunctional GTP cyclohydrolase II / 3,4-dihydroxy-2-butanone-4-phosphate synthase genes substantially similar to the *Arabidopsis* coding sequence (SEQ ID NO:13) from any other plant species.

Example 13: Expression and Purification of Recombinant Plant GTP Cyclohydrolase II / DHBP Synthase in *E. coli*.

To produce recombinant higher plant GTP cyclohydrolase II / 3,4-dihydroxy-2-butanone-4-phosphate synthase in *E. coli*, a translational fusion of the *Arabidopsis* GTP cyclohydrolase II / 3,4-dihydroxy-2-butanone-4-phosphate synthase cDNA (SEQ ID NO:13) to the 5' end of the thioredoxin gene (LaVallie et al., *Biotechnology* 11:187-193 (1992)) was created in pET-32a (Novagen, Inc., Madison, WI), using a two step PCR approach. Synthetic oligonucleotide primers DG-392a (SEQ ID NO:17) and DG-393a (SEQ ID NO:18) were used in a polymerase chain reaction to amplify a DNA fragment of 939-bp in length. The PCR product was digested with *NcoI* and *EcoRI*. The digestion products were separated on a low-gelling-temperature agarose gel and the fragments were excised. In parallel, plasmid pET32a was digested with *NcoI* and *EcoRI*. The digestion products were separated on a gel, and the pET32a vector was excised from the gel. The vector fragment was ligated to the PCR generated fragment, and the ligation products were transformed into competent *E. coli* XL1 Blue cells (Stratagene, La Jolla, CA).

Ampicillin-resistant colonies were selected, cultured, and their plasmid DNAs extracted. The structures of the plasmids were confirmed by sequencing with the chain

termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA). The recombinant plasmid with expected structure was designated pET32aGTP-1.

Synthetic oligonucleotide primers DG-390a (SEQ ID NO:19) and DG-391a (SEQ ID NO:20) were then used in a polymerase chain reaction to amplify a DNA fragment of 662-bp. The PCR product was digested with *NcoI*. The digestion products were separated on a low-gelling-temperature agarose gel and the fragments were excised. In parallel, plasmid pET32aGTP-1 was digested with *NcoI*. The digestion products were separated on a gel, and the pET32aGTP-1 vector was excised from the gel. The vector fragment was ligated to the PCR generated fragment, and the ligation products were transformed into competent *E. coli* XL1 Blue cells (Stratagene, La Jolla, CA).

Ampicillin-resistant colonies were selected, cultured, and their plasmid DNAs extracted. The structure of the plasmids were confirmed by sequencing with the chain termination method using dideoxy terminators labeled with fluorescent dyes (Applied Biosystems, Inc., Foster City, CA). The recombinant plasmid with expected structure was designated pET32aGTP-2.

Plasmid pET32aGTP-2 was transformed into competent *E. coli* BL21(DE3) cells, and recombinant protein was expressed and purified according to the manufacturer's instructions (pET System Manual, Novagen, Inc., Madison, WI). The resulting fusion proteins produced by this strain contained approximately 132 amino acids of *E. coli* thioredoxin protein, His-Tag, and thrombin cleavage site, followed by the presumptive mature coding sequence for *Arabidopsis* GTP cyclohydrolase II / 3,4-dihydroxy-2-butanone-4-phosphate synthase.

Example 14: *In vitro* Recombination of Riboflavin Biosynthesis Genes by DNA Snuffling

A plant riboflavin biosynthesis gene (e.g., SEQ ID NO:1 or SEQ ID NO:13) encoding a riboflavin biosynthesis protein (e.g., SEQ ID NO:2 or SEQ ID NO:14, respectively) is amplified by PCR. The resulting DNA fragment is digested by DNaseI treatment essentially as described (Stemmer et al. (1994) PNAS 91: 10747-10751) and the PCR primers are removed from the reaction mixture. A PCR reaction is carried out without primers and is followed by a PCR reaction with the primers, both as described (Stemmer et al. (1994) PNAS 91: 10747-10751). The resulting DNA fragments are cloned into pTRC99a

(Pharmacia, Cat no: 27-5007-01) and transformed into a dioxygenase mutant host, e.g. by electroporation using the Biorad Gene Pulser and the manufacturer's conditions. The transformed host is grown on medium that contains inhibitory concentrations of an inhibitor selected according to a method described above, and those colonies that grow in the presence of the inhibitor are selected. Colonies that grow in the presence of normally inhibitory concentrations of inhibitor are picked and purified by repeated restreaking. Their plasmids are purified and the DNA sequences of cDNA inserts from plasmids that pass this test are then determined.

In a similar reaction, PCR-amplified DNA fragments comprising a plant riboflavin biosynthesis gene of the invention encoding a riboflavin biosynthesis protein and PCR-amplified DNA fragments comprising a riboflavin biosynthesis gene from a different host are recombined *in vitro* and resulting variants with improved tolerance to the inhibitor are recovered as described above.

Example 15: *In vitro* Recombination of Riboflavin Biosynthesis Genes by Staggered Extension Process

A plant riboflavin biosynthesis gene (e.g., SEQ ID NO:1 or SEQ ID NO:13) encoding a riboflavin biosynthesis protein (e.g., SEQ ID NO:2 or SEQ ID NO:14, respectively) and a corresponding riboflavin biosynthesis gene from a different host are each cloned into the polylinker of a pBluescript vector. A PCR reaction is carried out essentially as described (Zhao et al. (1998) Nature Biotechnology 16: 258-261) using the "reverse primer" and the "M13 20 primer" (Stratagene Catalog). Amplified PCR fragments are digested with appropriate restriction enzymes and cloned into pTRC99a and mutated riboflavin biosynthesis genes are screened as described in Example 14.

Various modifications of the invention described herein will become apparent to those skilled in the art. Such modifications are intended to fall within the scope of the appended claims.

What Is Claimed Is:

1. A DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme involved in riboflavin biosynthesis, wherein the enzyme has lumazine synthase activity or bifunctional GTP cyclohydrolase II / DHBP synthase activity.
2. A DNA molecule according to claim 1, wherein the enzyme has lumazine synthase activity.
3. A DNA molecule according to claim 2, wherein the enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:2.
4. A DNA molecule according to claim 2, wherein the enzyme comprises the amino acid sequence set forth in SEQ ID NO:2.
5. A DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme having lumazine synthase activity, wherein said DNA molecule hybridizes to a DNA molecule according to claim 4 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C.
6. A DNA molecule according to claim 2, wherein said DNA molecule hybridizes to the coding sequence set forth in SEQ ID NO:1 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C.
7. A DNA molecule according to claim 2, wherein said DNA molecule comprises a 20 base pair nucleotide portion identical in sequence to a consecutive 20 base pair portion of the coding sequence set forth in SEQ ID NO:1.
8. A DNA molecule according to claim 2, comprising the coding sequence set forth in SEQ ID NO:1.

9. A DNA molecule according to claim 1, wherein the enzyme has bifunctional GTP cyclohydrolase II / DHBP synthase activity.
10. A DNA molecule according to claim 9, wherein the enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:14.
11. A DNA molecule according to claim 9, wherein the enzyme comprises the amino acid sequence set forth in SEQ ID NO:14.
12. A DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme having bifunctional GTP cyclohydrolase II / DHBP synthase activity, wherein said DNA molecule hybridizes to a DNA molecule according to claim 11 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C.
13. A DNA molecule according to claim 9, wherein said DNA molecule hybridizes to the coding sequence set forth in SEQ ID NO:13 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C.
14. A DNA molecule according to claim 9, wherein said DNA molecule comprises a 20 base pair nucleotide portion identical in sequence to a consecutive 20 base pair portion of the coding sequence set forth in SEQ ID NO:13.
15. A DNA molecule according to claim 9, comprising the coding sequence set forth in SEQ ID NO:13.
16. A chimeric gene comprising a promoter operatively linked to a DNA molecule according to claim 1.
17. A recombinant vector comprising a chimeric gene according to claim 16, wherein said vector is capable of being stably transformed into a host cell.

18. A host cell comprising a vector according to claim 17, wherein said host cell is capable of expressing the DNA molecule encoding an enzyme involved in riboflavin biosynthesis.
19. A host cell according to claim 18, wherein said host cell is selected from the group consisting of a bacterial cell, a yeast cell, and a plant cell.
20. A host cell according to claim 19, which is a bacterial cell.
21. A process for producing nucleotide sequences encoding gene products having altered lumazine synthase activity comprising:
- (a) shuffling a DNA molecule according to claim 2;
 - (b) expressing the resulting shuffled nucleotide sequences; and
 - (c) selecting for altered lumazine synthase activity as compared to the activity of an enzyme encoded by a DNA molecule according to claim 2.
22. The process of claim 21, wherein the nucleotide sequence is SEQ ID NO: 1.
23. A shuffled DNA molecule obtainable by the process of claim 22.
24. A shuffled DNA molecule according to claim 23, wherein said shuffled DNA molecule encodes an enzyme having enhanced tolerance to an inhibitor of lumazine synthase activity.
25. A chimeric gene comprising a promoter operatively linked to a shuffled DNA molecule according to claim 23.
26. A recombinant vector comprising a chimeric gene according to claim 25, wherein said vector is capable of being stably transformed into a host cell.
27. A host cell comprising a vector according to claim 26.
28. A host cell according to claim 27, wherein said host cell is selected from the group consisting of a bacterial cell, a yeast cell, and a plant cell.

29. A host cell according to claim 28, wherein said host cell is a plant cell.
30. A plant or seed comprising a plant cell according to claim 29.
31. A plant according to claim 30, wherein said plant is tolerant to an inhibitor of lumazine synthase activity.
32. A process for producing nucleotides sequences encoding gene products having altered bifunctional GTP cyclohydrolase II / DHBP synthase activity comprising:
- (a) shuffling a DNA molecule according to claim 9;
 - (b) expressing the resulting shuffled nucleotide sequences; and
 - (c) selecting for altered bifunctional GTP cyclohydrolase II / DHBP synthase activity as compared to the activity of an enzyme encoded by a DNA molecule according to claim 9.
33. The process of claim 32, wherein the nucleotide sequence is SEQ ID NO: 13.
34. A shuffled DNA molecule obtainable by the process of claim 33.
35. A shuffled DNA molecule according to claim 34, wherein said shuffled DNA molecule encodes an enzyme having enhanced tolerance to an inhibitor of bifunctional GTP cyclohydrolase II / DHBP synthase activity.
36. A chimeric gene comprising a promoter operatively linked to a shuffled DNA molecule according to claim 34.
37. A recombinant vector comprising a chimeric gene according to claim 36, wherein said vector is capable of being stably transformed into a host cell.
38. A host cell comprising a vector according to claim 37.
39. A host cell according to claim 38, wherein said host cell is selected from the group consisting of a bacterial cell, a yeast cell, and a plant cell.

40. A host cell according to claim 39, wherein said host cell is a plant cell.
41. A plant or seed comprising a plant cell according to claim 40.
42. A plant according to claim 41, wherein said plant is tolerant to an inhibitor of bifunctional GTP cyclohydrolase II / DHBP synthase activity.
43. An isolated plant enzyme involved in riboflavin biosynthesis, wherein said enzyme has lumazine synthase activity or bifunctional GTP cyclohydrolase II / DHBP synthase activity.
44. An enzyme according to claim 43, wherein said enzyme has lumazine synthase activity.
45. An enzyme according to claim 44, wherein said enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:2.
46. An enzyme according to claim 44, wherein said enzyme comprises the amino acid sequence set forth in SEQ ID NO:2.
47. An enzyme according to claim 43, wherein said enzyme has bifunctional GTP cyclohydrolase II / DHBP synthase activity.
48. An enzyme according to claim 47, wherein said enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:14.
49. An enzyme according to claim 47, wherein said enzyme comprises the amino acid sequence set forth in SEQ ID NO:14.
50. A method for screening a chemical for the ability to inhibit lumazine synthase activity, comprising the steps of:
- (a) combining an enzyme according to claim 44 in a first reaction mixture with 2,4-dioxy-5-amino-6-ribitylamino-pyrimidine and 3,4-dihydroxy-2-butanone phosphate under conditions in which the enzyme is capable of catalyzing the synthesis of lumazine;

- (b) combining the chemical and the enzyme in a second reaction mixture with 2,4-dioxy-5-amino-6-ribitylamino-pyrimidine and 3,4-dihydroxy-2-butanone phosphate under the same conditions as in the first reaction mixture;
- (c) determining the amounts of lumazine produced in the first and second reaction mixtures; and
- (d) comparing the amounts of lumazine produced in the first and second reaction mixtures;

wherein the chemical is capable of inhibiting the lumazine synthase activity of the enzyme if the amount of lumazine produced in the second reaction mixture is significantly less than the amount of lumazine produced in the first reaction mixture.

51. A method according to claim 50, wherein the first reaction mixture comprises 50 μ M 2,4-dioxy-5-amino-6-ribitylamino-pyrimidine, and 0.5 mM 3,4-dihydroxy-2-butanone phosphate.

52. A method according to claim 50, wherein the amounts of lumazine produced in the reaction mixtures are determined using a fluorimeter at an excitation wavelength of 407 nm.

53. A chemical identified by the screening method of claim 50.

54. A method for suppressing the growth of a plant, comprising applying to the plant the chemical of claim 53, whereby the chemical inhibits the activity of lumazine synthase in the plant.

55. A method for screening a chemical for the ability to inhibit bifunctional GTP cyclohydrolase II / DHBP synthase activity, comprising the steps of:

- (a) combining an enzyme according to claim 47 in a first reaction mixture with GTP or ribulose-5-phosphate under conditions in which the enzyme is capable of catalyzing the synthesis of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate, respectively;
- (b) combining the chemical and the enzyme in a second reaction mixture with GTP or ribulose-5-phosphate under the same conditions as in the first reaction mixture;

- (c) determining the amounts of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate produced in the first and second reaction mixtures; and
- (d) comparing the amounts of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate produced in the first and second reaction mixtures;

wherein the chemical is capable of inhibiting the bifunctional GTP cyclohydrolase II / DHBP synthase activity of the enzyme if the amount of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate produced in the second reaction mixture is significantly less than the amount of 2,5-diamino-4-oxy-6-ribosylamino-pyrimidine-5'-phosphate or 3,4-dihydroxy-2-butanone phosphate produced in the first reaction mixture.

56. A chemical identified by the screening method of claim 55.

57. A method for suppressing the growth of a plant, comprising applying to the plant the chemical of claim 56, whereby the chemical inhibits the activity of GTP cyclohydrolase II / DHBP synthase in the plant.

58. A plant, plant cell, plant seed, or plant tissue comprising a DNA molecule comprising a nucleotide sequence isolated from a plant that encodes an enzyme involved in riboflavin biosynthesis, wherein the enzyme has lumazine synthase activity or bifunctional GTP cyclohydrolase II / DHBP synthase activity, and wherein the DNA molecule confers upon said plant, plant cell, plant seed, or plant tissue tolerance to a herbicide in amounts that normally inhibit riboflavin biosynthesis.

59. A plant, plant cell, plant seed, or plant tissue according to claim 58, wherein the enzyme has lumazine synthase activity, and wherein the DNA molecule confers upon the plant, plant cell, plant seed, or plant tissue tolerance to a herbicide in amounts that inhibit naturally occurring lumazine synthase activity.

60. A plant, plant cell, plant seed, or plant tissue according to claim 59, wherein the enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:2.

61. A plant, plant cell, plant seed, or plant tissue according to claim 59, wherein the DNA molecule hybridizes to the coding sequence set forth in SEQ ID NO:1 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C.
62. A plant, plant cell, plant seed, or plant tissue according to claim 58, wherein the enzyme has bifunctional GTP cyclohydrolase II / DHBP synthase activity, and wherein the DNA molecule confers upon the plant, plant cell, plant seed, or plant tissue tolerance to a herbicide in amounts that inhibit naturally occurring bifunctional GTP cyclohydrolase II / DHBP synthase activity.
63. A plant, plant cell, plant seed, or plant tissue according to claim 62, wherein the enzyme comprises an amino acid sequence substantially similar to the amino acid sequence set forth in SEQ ID NO:14.
64. A plant, plant cell, plant seed, or plant tissue according to claim 62, wherein the DNA molecule hybridizes to the coding sequence set forth in SEQ ID NO:13 under the following conditions: hybridization at 7% sodium dodecyl sulfate (SDS), 0.5 M NaPO₄ pH 7.0, 1 mM EDTA at 50°C; wash with 2X SSC, 1% SDS, at 50°C.
65. A method for selectively suppressing the growth of weeds in a field containing a crop of planted crop seeds or plants, comprising the steps of:
- (a) planting herbicide tolerant crops or crop seeds, which are plants or plant seeds according to claim 59; and
 - (b) applying to the crops or crop seeds and the weeds in the field a herbicide in amounts that inhibit naturally occurring lumazine synthase activity, wherein the herbicide suppresses the growth of the weeds without significantly suppressing the growth of the crops.
66. A method for selectively suppressing the growth of weeds in a field containing a crop of planted crop seeds or plants, comprising the steps of:

- (a) planting herbicide tolerant crops or crop seeds, which are plants or plant seeds according to claim 39; and
- (b) applying to the crops or crop seeds and the weeds in the field a herbicide in amounts that inhibit naturally occurring bifunctional GTP cyclohydrolase II / DHBP synthase activity, wherein the herbicide suppresses the growth of the weeds without significantly suppressing the growth of the crops.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:

- (A) NAME: Novartis AG
- (B) STREET: Schwarzwaldallee 215
- (C) CITY: Basel
- (E) COUNTRY: Switzerland
- (F) POSTAL CODE (ZIP): 4058
- (G) TELEPHONE: +41 61 324 11 11
- (H) TELEFAX: + 41 61 322 75 32

(ii) TITLE OF INVENTION: RIBOFLAVIN BIOSYNTHESIS GENES FROM PLANTS AND USES THEREOF

(iii) NUMBER OF SEQUENCES: 20

(iv) COMPUTER READABLE FORM:

- (A) MEDIUM TYPE: Floppy disk
- (B) COMPUTER: IBM PC compatible
- (C) OPERATING SYSTEM: PC-DOS/MS-DOS
- (D) SOFTWARE: PatentIn Release #1.0, Version #1.30

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 991 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:

- (A) NAME/KEY: CDS
- (B) LOCATION: 35..718

(D) OTHER INFORMATION: /product= "lumazine synthase"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```

AGAGAACCGT CTCTAAACT CCGACGAACG AAAA ATG AAG TCA TTA GCT TCG
52
Met Lys Ser Leu Ala Ser
1 5

CCG CCG TGT CTC CGC CTG ATA CCG ACG GCA CAC CGT CAG CTC AAT TCG
100
Pro Pro Cys Leu Arg Leu Ile Pro Thr Ala His Arg Gln Leu Asn Ser
10 15 20

CGT CAA TCT TCC TCC GCC TGT TAT ATA CAC GGT GGC TCT TCT GTG AAC
148
Arg Gln Ser Ser Ser Ala Cys Tyr Ile His Gly Gly Ser Ser Val Asn
25 30 35

AAA TCC AAT AAT CTC TCA TTC TCC TCA TCC ACA TCC GGA TTT GCG TCA
196
Lys Ser Asn Asn Leu Ser Phe Ser Ser Ser Thr Ser Gly Phe Ala Ser
40 45 50

CCA CTA GCT GTA GAG AAG GAA TTA CGC TCT TCA TTC GTA CAG ACG GCT
244
Pro Leu Ala Val Glu Lys Glu Leu Arg Ser Ser Phe Val Gln Thr Ala
55 60 65 70

GCT GTT CGC CAT GTT ACG GGG TCT CTT ATC AGA GGC GAA GGT CTT AGA
292
Ala Val Arg His Val Thr Gly Ser Leu Ile Arg Gly Glu Gly Leu Arg
75 80 85

TTC GCC ATC GTG GTA GCT CGT TTC AAT GAG GTT GTG ACT AAG TTG CTT
340
Phe Ala Ile Val Val Ala Arg Phe Asn Glu Val Val Thr Lys Leu Leu
90 95 100

```

TTG GAA GGA GCG ATT GAG ACT TTC AAG AAG TAT TCA GTC AGA GAA GAA
388

Leu Glu Gly Ala Ile Glu Thr Phe Lys Lys Tyr Ser Val Arg Glu Glu
105 110 115

GAC ATT GAA GTT ATT TGG GTT CCT GGC AGC TTT GAA ATT GGT GTT GTT
436

Asp Ile Glu Val Ile Trp Val Pro Gly Ser Phe Glu Ile Gly Val Val
120 125 130

GCA CAA AAT CTT GGG AAA TCG GGA AAA TTT CAT GCT GTT TTA TGT ATC
484

Ala Gln Asn Leu Gly Lys Ser Gly Lys Phe His Ala Val Leu Cys Ile
135 140 145 150

GGC GCT GTG ATA AGA GGA GAT ACC ACA CAT TAT GAT GCT GTT GCC AAC
532

Gly Ala Val Ile Arg Gly Asp Thr Thr His Tyr Asp Ala Val Ala Asn
155 160 165

TCT GCT GCG TCT GGA GTA CTT TCT GCT AGC ATA AAT TCA GGC GTT CCA
580

Ser Ala Ala Ser Gly Val Leu Ser Ala Ser Ile Asn Ser Gly Val Pro
170 175 180

TGC ATA TTT GGT GTA CTG ACT TGC GAG GAC ATG GAT CAG GCT CTG AAT
628

Cys Ile Phe Gly Val Leu Thr Cys Glu Asp Met Asp Gln Ala Leu Asn
185 190 195

CGA TCT GGT GGC AAA GCC GGC AAT AAG GGA GCT GAA ACT GCT TTG ACG
676

Arg Ser Gly Gly Lys Ala Gly Asn Lys Gly Ala Glu Thr Ala Leu Thr
200 205 210

GCG CTC GAA ATG GCG TCG TTG TTT GAG CAC CAC CTG AAA TAG
718

Ala Leu Glu Met Ala Ser Leu Phe Glu His His Leu Lys *

215 220 225
 CTCGGCTCGT TCGATGGATG AACATGATCA CGTATGAGAA CCTCTTGATG TTGTCCCATT
 778
 TGGTTACAAT CCAGTCTCTG AAATTGTTTG TACCTCAAAG ATTGTCCAAA TGTTTTACCC
 838
 TTGGTTACCA AATCAATTAA ACGCTTTTGT AAGCTTCTGG CCTTGTTTTT TTTTTTTGAA
 898
 TCGTATGATA ATAATAATTC CTCCGAATTT TGGGGTCTTT CTGTACTAAT CAAAAATGTG
 958
 ATCTTCTTTG TTGTAAAAAA AAAAAAAAAA AAA
 991

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 228 amino acids
- (B) TYPE: amino acid
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

Met Lys Ser Leu Ala Ser Pro Pro Cys Leu Arg Leu Ile Pro Thr Ala
 1 5 10 15

His Arg Gln Leu Asn Ser Arg Gln Ser Ser Ser Ala Cys Tyr Ile His
 20 25 30

Gly Gly Ser Ser Val Asn Lys Ser Asn Asn Leu Ser Phe Ser Ser Ser
 35 40 45

Thr Ser Gly Phe Ala Ser Pro Leu Ala Val Glu Lys Glu Leu Arg Ser

50 55 60
 Ser Phe Val Gln Thr Ala Ala Val Arg His Val Thr Gly Ser Leu Ile
 65 70 75 80
 Arg Gly Glu Gly Leu Arg Phe Ala Ile Val Val Ala Arg Phe Asn Glu
 85 90 95
 Val Val Thr Lys Leu Leu Leu Glu Gly Ala Ile Glu Thr Phe Lys Lys
 100 105 110
 Tyr Ser Val Arg Glu Glu Asp Ile Glu Val Ile Trp Val Pro Gly Ser
 115 120 125
 Phe Glu Ile Gly Val Val Ala Gln Asn Leu Gly Lys Ser Gly Lys Phe
 130 135 140
 His Ala Val Leu Cys Ile Gly Ala Val Ile Arg Gly Asp Thr Thr His
 145 150 155 160
 Tyr Asp Ala Val Ala Asn Ser Ala Ala Ser Gly Val Leu Ser Ala Ser
 165 170 175
 Ile Asn Ser Gly Val Pro Cys Ile Phe Gly Val Leu Thr Cys Glu Asp
 180 185 190
 Met Asp Gln Ala Leu Asn Arg Ser Gly Gly Lys Ala Gly Asn Lys Gly
 195 200 205
 Ala Glu Thr Ala Leu Thr Ala Leu Glu Met Ala Ser Leu Phe Glu His
 210 215 220
 His Leu Lys *
 225

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 base pairs

- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: other nucleic acid
 - (A) DESCRIPTION: /desc = "DG-63"

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

ATTTTGTAAC CAAGGG

16

- (2) INFORMATION FOR SEQ ID NO:4:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 16 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: other nucleic acid
 - (A) DESCRIPTION: /desc = "DG-65"

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

GGCAATAAGG GAGCTG

16

- (2) INFORMATION FOR SEQ ID NO:5:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 22 base pairs
 - (B) TYPE: nucleic acid

- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: other nucleic acid
 - (A) DESCRIPTION: /desc = "JG-L"

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

GTACCTCGAG TCTAGACTCG AG
22

- (2) INFORMATION FOR SEQ ID NO:6:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 27 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: other nucleic acid
 - (A) DESCRIPTION: /desc = "RS-1"

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

AGCTACCATG GGAGGTTCTC ATACGTG
27

- (2) INFORMATION FOR SEQ ID NO:7:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 27 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

(A) DESCRIPTION: /desc = "RS-2"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

AGCTAGAGCT CACGAGAGAA CCGTCTC

27

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 11 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS: not relevant

(D) TOPOLOGY: not relevant

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

Cys	Ile	Gly	Ala	Val	Ile	Arg	Gly	Asp	Thr	Thr
1				5					10	

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS: not relevant

(D) TOPOLOGY: not relevant

(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

Met Lys Ala Gly Asn Lys Gly Ala Glu Thr Ala Leu Thr Ala Leu Glu
1 5 10 15

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

(A) DESCRIPTION: /desc = "DG-252"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GATCCCATGG CTAAGTCATT AGCTTCGCCG
30

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 27 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: other nucleic acid
 - (A) DESCRIPTION: /desc = "DG-253"

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

ATCGCCATGG CTGTTGCGCA TGTTACG

27

- (2) INFORMATION FOR SEQ ID NO:12:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 31 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: other nucleic acid
 - (A) DESCRIPTION: /desc = "DG-254"

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

CAGTGAATTC CTAGAGCTAT TTCAGGTGGT G

31

- (2) INFORMATION FOR SEQ ID NO:13:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 1665 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: cDNA

(ix) FEATURE:

(A) NAME/KEY: CDS
(B) LOCATION: 2..1432
(D) OTHER INFORMATION: /product= "bifunctional GTP
cyclohydrolase II / DHBP synthase"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

C TCA TTC ACC AAC GGA AAC ACT CCT CTC TCA AAT GGG TCT CTC ATT
46

Ser Phe Thr Asn Gly Asn Thr Pro Leu Ser Asn Gly Ser Leu Ile
1 5 10 15

GAT GAT CGG ACC GAA GAG CCA TTA GAG GCT GAT TCG GTT TCA CTT GGA
94

Asp Asp Arg Thr Glu Glu Pro Leu Glu Ala Asp Ser Val Ser Leu Gly
20 25 30

ACA CTT GCT GCT GAT TCT GCT CCT GCA CCA GCC AAT GGT TTT GTT GCT
142

Thr Leu Ala Ala Asp Ser Ala Pro Ala Pro Ala Asn Gly Phe Val Ala
35 40 45

GAA GAT GAT GAC TTT GAG TTG GAT TTA CCA ACT CCT GGT TTC TCT TCT
190

Glu Asp Asp Asp Phe Glu Leu Asp Leu Pro Thr Pro Gly Phe Ser Ser
50 55 60

ATC CCT GAG GCC ATT GAA GAT ATA CGC CAA GGA AAG CTT GTG GTG GTT
238

Ile Pro Glu Ala Ile Glu Asp Ile Arg Gln Gly Lys Leu Val Val Val
65 70 75

GTG GAT GAT GAA GAT AGG GAA AAT GAA GGG GAT TTG GTG ATG GCT GCT
286

Val Asp Asp Glu Asp Arg Glu Asn Glu Gly Asp Leu Val Met Ala Ala

80 85 90 95
 CAG TTA GCA ACA CCT GAA GCT ATG GCT TTT ATT GTG AGA CAT GGA ACT
 334
 Gln Leu Ala Thr Pro Glu Ala Met Ala Phe Ile Val Arg His Gly Thr
 100 105 110
 GGG ATA GTT TGT GTG AGC ATG AAA GAA GAT GAT CTC GAG AGG TTG CAC
 382
 Gly Ile Val Cys Val Ser Met Lys Glu Asp Asp Leu Glu Arg Leu His
 115 120 125
 CTT CCT CTA ATG GTG AAT CAG AAG GAA AAC GAA GAA AAG CTC TCT ACT
 430
 Leu Pro Leu Met Val Asn Gln Lys Glu Asn Glu Glu Lys Leu Ser Thr
 130 135 140
 GCA TTT ACA GTG ACT GTG GAT GCA AAA CAT GGC ACA ACA ACG GGA GTA
 478
 Ala Phe Thr Val Thr Val Asp Ala Lys His Gly Thr Thr Thr Gly Val
 145 150 155
 TCA GCT CGT GAC AGG GCA ACA ACC ATA TTG TCT CTT GCA TCA AGA GAT
 526
 Ser Ala Arg Asp Arg Ala Thr Thr Ile Leu Ser Leu Ala Ser Arg Asp
 160 165 170 175
 TCA AAG CCT GAG GAT TTC AAT CGT CCA GGT CAT ATC TTC CCA CTG AAG
 574
 Ser Lys Pro Glu Asp Phe Asn Arg Pro Gly His Ile Phe Pro Leu Lys
 180 185 190
 TAT CGG GAA GGT GGG GTT CTG AAA AGG GCT GGA CAC ACT GAA GCA TCT
 622
 Tyr Arg Glu Gly Gly Val Leu Lys Arg Ala Gly His Thr Glu Ala Ser
 195 200 205
 GTT GAT CTC ACT GTT TTA GCT GGA CTG GAT CCT GTT GGA GTA CTT TGT
 670

Val Asp Leu Thr Val Leu Ala Gly Leu Asp Pro Val Gly Val Leu Cys
 210 215 220

GAA ATT GTT GAT GAT GAT GGT TCC ATG GCT AGA TTA CCA AAA CTT CGT
 718

Glu Ile Val Asp Asp Asp Gly Ser Met Ala Arg Leu Pro Lys Leu Arg
 225 230 235

GAA TTT GCC GCC GAG AAC AAC CTG AAA GTT GTT TCC ATC GCA GAT TTG
 766

Glu Phe Ala Ala Glu Asn Asn Leu Lys Val Val Ser Ile Ala Asp Leu
 240 245 250 255

ATC AGG TAT AGA AGA AAG AGA GAT AAA TTA GTG GAA CGT GCT TCT GCG
 814

Ile Arg Tyr Arg Arg Lys Arg Asp Lys Leu Val Glu Arg Ala Ser Ala
 260 265 270

GCT CGG ATC CCA ACA ATG TGG GGA CCT TTC ACT GCT TAC TGC TAT AGG
 862

Ala Arg Ile Pro Thr Met Trp Gly Pro Phe Thr Ala Tyr Cys Tyr Arg
 275 280 285

TCC ATA TTA GAC GGA ATA GAG CAC ATA GCA ATG GTT AAG GGT GAG ATT
 910

Ser Ile Leu Asp Gly Ile Glu His Ile Ala Met Val Lys Gly Glu Ile
 290 295 300

GGT GAC GGT CAA GAC ATT CTC GTG AGG GTT CAT TCT GAA TGT CTA ACA
 958

Gly Asp Gly Gln Asp Ile Leu Val Arg Val His Ser Glu Cys Leu Thr
 305 310 315

GGG GAC ATA TTT GGG TCT GCA AGG TGT GAT TGC GGG AAC CAG CTA GCA
 1006

Gly Asp Ile Phe Gly Ser Ala Arg Cys Asp Cys Gly Asn Gln Leu Ala
 320 325 330 335

CTC TCG ATG CAG CAG ATC GAG GCT ACT GGT CGC GGT GTG CTG GTT TAC
 1054

Leu Ser Met Gln Gln Ile Glu Ala Thr Gly Arg Gly Val Leu Val Tyr
340 345 350

CTA CGT GGA CAT GAA GGA AGA GGG ATC GGT TTA GGA CAC AAG CTT CGA
1102

Leu Arg Gly His Glu Gly Arg Gly Ile Gly Leu Gly His Lys Leu Arg
355 360 365

GCT TAC AAT CTG CAA GAT GCT GGT CGA GAC ACG GTT GAA GCT AAT GAG
1150

Ala Tyr Asn Leu Gln Asp Ala Gly Arg Asp Thr Val Glu Ala Asn Glu
370 375 380

GAA TTA GGA CTT CCT GTT GAT TCT AGA GAG TAT GGA ATT GGT GCA CAG
1198

Glu Leu Gly Leu Pro Val Asp Ser Arg Glu Tyr Gly Ile Gly Ala Gln
385 390 395

ATA ATA AGG GAT TTA GGT GTT AGG ACA ATG AAG CTG ATG ACA AAT AAT
1246

Ile Ile Arg Asp Leu Gly Val Arg Thr Met Lys Leu Met Thr Asn Asn
400 405 410 415

CCC CCA AAG TAT GTT GGT TTG AAG GGA TAT GGA TTA GCC ATT GTT GGG
1294

Pro Pro Lys Tyr Val Gly Leu Lys Gly Tyr Gly Leu Ala Ile Val Gly
420 425 430

AGA GTC CCT CTA TTG AGT CTT ATC ACG AAG GAG AAT AAG AGA TAT CTG
1342

Arg Val Pro Leu Leu Ser Leu Ile Thr Lys Glu Asn Lys Arg Tyr Leu
435 440 445

GAG ACA AAG CGG ACC AAG ATG GGT CAC ATG TAT GGC TTG AAG TTC AAA
1390

Glu Thr Lys Arg Thr Lys Met Gly His Met Tyr Gly Leu Lys Phe Lys
450 455 460

GGG GAT GTT GTG GAG AAG ATT GAG TCT GAA TCT GAG TCC TAA
1432

Gly Asp Val Val Glu Lys Ile Glu Ser Glu Ser Glu Ser *
 465 470 475

GCTTAAAAAC CAGGACGAAC CGAATGGAAT CAAGAACTAT AGATATAATA CTTCCCAAAA
 1492

AACAAGGAAA GAAATTGACA CAGAAGAAGA GGAAAAAGAC ATTTGATCTG TCTGAGAAAC
 1552

TTGATTAGAT TGGTTTATGT TCTAATCTAA TCTGATTGA TTTTTTTTAA TTTTGTCTAC
 1612

GATTCTTGAG TTACGAAATG TTCATCATTT GTTAAAAAAA AAAAAAAAAA AAA
 1665

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 477 amino acids
- (B) TYPE: amino acid
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

Ser Phe Thr Asn Gly Asn Thr Pro Leu Ser Asn Gly Ser Leu Ile Asp
 1 5 10 15

Asp Arg Thr Glu Glu Pro Leu Glu Ala Asp Ser Val Ser Leu Gly Thr
 20 25 30

Leu Ala Ala Asp Ser Ala Pro Ala Pro Ala Asn Gly Phe Val Ala Glu
 35 40 45

Asp Asp Asp Phe Glu Leu Asp Leu Pro Thr Pro Gly Phe Ser Ser Ile
 50 55 60

Pro Glu Ala Ile Glu Asp Ile Arg Gln Gly Lys Leu Val Val Val Val
 65 70 75 80

Asp Asp Glu Asp Arg Glu Asn Glu Gly Asp Leu Val Met Ala Ala Gln
 85 90 95

Leu Ala Thr Pro Glu Ala Met Ala Phe Ile Val Arg His Gly Thr Gly
 100 105 110

Ile Val Cys Val Ser Met Lys Glu Asp Asp Leu Glu Arg Leu His Leu
 115 120 125

Pro Leu Met Val Asn Gln Lys Glu Asn Glu Glu Lys Leu Ser Thr Ala
 130 135 140

Phe Thr Val Thr Val Asp Ala Lys His Gly Thr Thr Thr Gly Val Ser
 145 150 155 160

Ala Arg Asp Arg Ala Thr Thr Ile Leu Ser Leu Ala Ser Arg Asp Ser
 165 170 175

Lys Pro Glu Asp Phe Asn Arg Pro Gly His Ile Phe Pro Leu Lys Tyr
 180 185 190

Arg Glu Gly Gly Val Leu Lys Arg Ala Gly His Thr Glu Ala Ser Val
 195 200 205

Asp Leu Thr Val Leu Ala Gly Leu Asp Pro Val Gly Val Leu Cys Glu
 210 215 220

Ile Val Asp Asp Asp Gly Ser Met Ala Arg Leu Pro Lys Leu Arg Glu
 225 230 235 240

Phe Ala Ala Glu Asn Asn Leu Lys Val Val Ser Ile Ala Asp Leu Ile
 245 250 255

Arg Tyr Arg Arg Lys Arg Asp Lys Leu Val Glu Arg Ala Ser Ala Ala
 260 265 270

Arg Ile Pro Thr Met Trp Gly Pro Phe Thr Ala Tyr Cys Tyr Arg Ser
 275 280 285

Ile Leu Asp Gly Ile Glu His Ile Ala Met Val Lys Gly Glu Ile Gly
 290 295 300

Asp Gly Gln Asp Ile Leu Val Arg Val His Ser Glu Cys Leu Thr Gly
 305 310 315 320

Asp Ile Phe Gly Ser Ala Arg Cys Asp Cys Gly Asn Gln Leu Ala Leu
 325 330 335

Ser Met Gln Gln Ile Glu Ala Thr Gly Arg Gly Val Leu Val Tyr Leu
 340 345 350

Arg Gly His Glu Gly Arg Gly Ile Gly Leu Gly His Lys Leu Arg Ala
 355 360 365

Tyr Asn Leu Gln Asp Ala Gly Arg Asp Thr Val Glu Ala Asn Glu Glu
 370 375 380

Leu Gly Leu Pro Val Asp Ser Arg Glu Tyr Gly Ile Gly Ala Gln Ile
 385 390 395 400

Ile Arg Asp Leu Gly Val Arg Thr Met Lys Leu Met Thr Asn Asn Pro
 405 410 415

Pro Lys Tyr Val Gly Leu Lys Gly Tyr Gly Leu Ala Ile Val Gly Arg
 420 425 430

Val Pro Leu Leu Ser Leu Ile Thr Lys Glu Asn Lys Arg Tyr Leu Glu
 435 440 445

Thr Lys Arg Thr Lys Met Gly His Met Tyr Gly Leu Lys Phe Lys Gly
 450 455 460

Asp Val Val Glu Lys Ile Glu Ser Glu Ser Glu Ser *
 465 470 475

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 16 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

- (A) DESCRIPTION: /desc = "DG-67"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GCTAATGAGG AATTAG

16

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 16 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

- (A) DESCRIPTION: /desc = "DG-69"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

TGATTCCATT CGGTTC

16

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 18 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

- (A) DESCRIPTION: /desc = "DG-392a"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

TGTCTCTTGC ATCAAGAG

18

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 33 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

- (A) DESCRIPTION: /desc = "DG-393a"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

CAGTGAATTC TTAAGCTTAG GACTCAGATT CAG

33

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 25 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

- (A) DESCRIPTION: /desc = "DG-390a"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

GATCCCATGG GTTCTCTTC TATCG

25

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 18 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

- (A) DESCRIPTION: /desc = "DG-391a"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

CCGAGCCGCA GAAGCACG

18

121

The DNA substrate molecules to be digested can either be from *in vivo* replicated DNA, such as a plasmid preparation, or from PCR amplified nucleic acid fragments harboring the restriction enzyme recognition sites of interest, preferably near the ends of the fragment. Typically, at least two variants of a gene of interest, each having one or more mutations, are digested with at least one restriction enzyme determined to cut within the nucleic acid sequence of interest. The restriction fragments are then joined with DNA ligase to generate full length genes having shuffled regions. The number of regions shuffled will depend on the number of cuts within the nucleic acid sequence of interest. The shuffled molecules can be introduced into cells as described above and screened or selected for a desired property as described herein. Nucleic acid can then be isolated from pools (libraries), or clones having desired properties and subjected to the same procedure until a desired degree of improvement is obtained.

In some embodiments, at least one DNA substrate molecule or fragment thereof is isolated and subjected to mutagenesis. In some embodiments, the pool or library of religated restriction fragments are subjected to mutagenesis before the digestion-ligation process is repeated. "Mutagenesis" as used herein includes such techniques known in the art as PCR mutagenesis, oligonucleotide-directed mutagenesis, site-directed mutagenesis, *etc.*, and recursive sequence recombination by any of the techniques described herein.

20 2. Reassembly PCR

A further technique for recombining mutations in a nucleic acid sequence utilizes "reassembly PCR." This method can be used to assemble multiple segments that have been separately evolved into a full length nucleic acid template such as a gene. This technique is performed when a pool of advantageous mutants is known from previous work or has been identified by screening mutants that may have been created by any mutagenesis technique known in the art, such as PCR mutagenesis, cassette mutagenesis, doped oligo mutagenesis, chemical mutagenesis, or propagation of the DNA template *in vivo* in mutator strains. Boundaries defining segments of a nucleic acid sequence of interest preferably lie in intergenic regions, introns, or areas of a gene not likely to have mutations of interest. Preferably, oligonucleotide primers (oligos) are synthesized for PCR amplification of segments of the nucleic acid sequence of interest, such that the sequences of the oligonucleotides overlap the junctions of two segments. The overlap region is typically about 10 to 100 nucleotides in length. Each of the segments is amplified with a set of such

primers. The PCR products are then "reassembled" according to assembly protocols such as those discussed herein to assemble randomly fragmented genes. In brief, in an assembly protocol the PCR products are first purified away from the primers, by, for example, gel electrophoresis or size exclusion chromatography. Purified products are mixed together and subjected to about 1-10 cycles of denaturing, reannealing, and extension in the presence of polymerase and deoxynucleoside triphosphates (dNTP's) and appropriate buffer salts in the absence of additional primers ("self-priming"). Subsequent PCR with primers flanking the gene are used to amplify the yield of the fully reassembled and shuffled genes.

In some embodiments, the resulting reassembled genes are subjected to mutagenesis before the process is repeated.

In a further embodiment, the PCR primers for amplification of segments of the nucleic acid sequence of interest are used to introduce variation into the gene of interest as follows. Mutations at sites of interest in a nucleic acid sequence are identified by screening or selection, by sequencing homologues of the nucleic acid sequence, and so on.

Oligonucleotide PCR primers are then synthesized which encode wild type or mutant information at sites of interest. These primers are then used in PCR mutagenesis to generate libraries of full length genes encoding permutations of wild type and mutant information at the designated positions. This technique is typically advantageous in cases where the screening or selection process is expensive, cumbersome, or impractical relative to the cost of sequencing the genes of mutants of interest and synthesizing mutagenic oligonucleotides.

3. Site Directed Mutagenesis (SDM) with Oligonucleotides Encoding Homologue Mutations Followed by Shuffling

In some embodiments of the invention, sequence information from one or more substrate sequences is added to a given "parental" sequence of interest, with subsequent recombination between rounds of screening or selection. Typically, this is done with site-directed mutagenesis performed by techniques well known in the art (e.g., Berger, Ausubel and Sambrook, *supra.*) with one substrate as template and oligonucleotides encoding single or multiple mutations from other substrate sequences, e.g. homologous genes. After screening or selection for an improved phenotype of interest, the selected recombinant(s) can be further evolved using RSR techniques described herein. After screening or selection, site-directed mutagenesis can be done again with another collection

of oligonucleotides encoding homologue mutations, and the above process repeated until the desired properties are obtained.

When the difference between two homologues is one or more single point mutations in a codon, degenerate oligonucleotides can be used that encode the sequences in both homologues. One oligonucleotide can include many such degenerate codons and still allow one to exhaustively search all permutations over that block of sequence.

When the homologue sequence space is very large, it can be advantageous to restrict the search to certain variants. Thus, for example, computer modeling tools (Lathrop *et al.*, *J. Mol. Biol.* 255:641-665 (1996)) can be used to model each homologue mutation onto the target protein and discard any mutations that are predicted to grossly disrupt structure and function.

4. *In vitro* DNA Shuffling Formats

In one embodiment for shuffling DNA sequences *in vitro*, the initial substrates for recombination are a pool of related sequences, *e.g.*, different variant forms, as homologs from different individuals, strains, or species of an organism, or related sequences from the same organism, as allelic variations. The sequences can be DNA or RNA and can be of various lengths depending on the size of the gene or DNA fragment to be recombined or reassembled. Preferably the sequences are from 50 base pairs (bp) to 50 kilobases (kb).

The pool of related substrates are converted into overlapping fragments, *e.g.*, from about 5 bp to 5 kb or more. Often, for example, the size of the fragments is from about 10 bp to 1000 bp, and sometimes the size of the DNA fragments is from about 100 bp to 500 bp. The conversion can be effected by a number of different methods, such as DNase I or RNase digestion, random shearing or partial restriction enzyme digestion. For discussions of protocols for the isolation, manipulation, enzymatic digestion, and the like of nucleic acids, see, for example, Sambrook *et al.* and Ausubel, both *supra*. The concentration of nucleic acid fragments of a particular length and sequence is often less than 0.1 % or 1% by weight of the total nucleic acid. The number of different specific nucleic acid fragments in the mixture is usually at least about 100, 500 or 1000.

The mixed population of nucleic acid fragments are converted to at least partially single-stranded form using a variety of techniques, including, for example, heating, chemical denaturation, use of DNA binding proteins, and the like. Conversion can be effected by heating to about 80 °C to 100 °C, more preferably from 90 °C to 96 °C, to form

single-stranded nucleic acid fragments and then reannealing. Conversion can also be effected by treatment with single-stranded DNA binding protein (see Wold, *Annu. Rev. Biochem.* 66:61-92 (1997)) or *recA* protein (see, e.g., Kiianitsa, *Proc. Natl. Acad. Sci. U S A* 94:7837-7840 (1997)). Single-stranded nucleic acid fragments having regions of sequence identity with other single-stranded nucleic acid fragments can then be reannealed by cooling to 20 °C to 75 °C, and preferably from 40 °C to 65 °C. Renaturation can be accelerated by the addition of polyethylene glycol (PEG), other volume-excluding reagents or salt. The salt concentration is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%. The fragments that reanneal can be from different substrates. The annealed nucleic acid fragments are incubated in the presence of a nucleic acid polymerase, such as Taq or Klenow, and dNTP's (*i.e.* dATP, dCTP, dGTP and dTTP). If regions of sequence identity are large, Taq polymerase can be used with an annealing temperature of between 45-65 °C. If the areas of identity are small, Klenow polymerase can be used with an annealing temperature of between 20-30 °C. The polymerase can be added to the random nucleic acid fragments prior to annealing, simultaneously with annealing or after annealing.

The process of denaturation, renaturation and incubation in the presence of polymerase of overlapping fragments to generate a collection of polynucleotides containing different permutations of fragments is sometimes referred to as shuffling of the nucleic acid *in vitro*. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 100 times, more preferably the sequence is repeated from 10 to 40 times. The resulting nucleic acids are a family of double-stranded polynucleotides of from about 50 bp to about 100 kb, preferably from 500 bp to 50 kb. The population represents variants of the starting substrates showing substantial sequence identity thereto but also diverging at several positions. The population has many more members than the starting substrates. The population of fragments resulting from shuffling is used to transform host cells, optionally after cloning into a vector.

In one embodiment utilizing *in vitro* shuffling, subsequences of recombination substrates can be generated by amplifying the full-length sequences under conditions which produce a substantial fraction, typically at least 20 percent or more, of incompletely extended amplification products. Another embodiment uses random primers to prime the entire template DNA to generate less than full length amplification products. The

amplification products, including the incompletely extended amplification products are denatured and subjected to at least one additional cycle of reannealing and amplification. This variation, in which at least one cycle of reannealing and amplification provides a substantial fraction of incompletely extended products, is termed "stuttering." In the subsequent amplification round, the partially extended (less than full length) products reanneal to and prime extension on different sequence-related template species. In another embodiment, the conversion of substrates to fragments can be effected by partial PCR amplification of substrates.

In another embodiment, a mixture of fragments is spiked with one or more oligonucleotides. The oligonucleotides can be designed to include precharacterized mutations of a wildtype sequence, or sites of natural variations between individuals or species. The oligonucleotides also include sufficient sequence or structural homology flanking such mutations or variations to allow annealing with the wildtype fragments. Annealing temperatures can be adjusted depending on the length of homology.

In a further embodiment, recombination occurs in at least one cycle by template switching, such as when a DNA fragment derived from one template primes on the homologous position of a related but different template. Template switching can be induced by addition of *recA* (see, Kiiianitsa *supra* (1997)), *rad51* (see, Namsaraev, *Mol. Cell. Biol.* 17:5359-5368 (1997)), *rad55* (see, Clever, *EMBO J.* 16:2535-2544 (1997)), *rad57* (see, Sung, *Genes Dev.* 11:1111-1121 (1997)) or other polymerases (e.g., viral polymerases, reverse transcriptase) to the amplification mixture. Template switching can also be increased by increasing the DNA template concentration.

Another embodiment utilizes at least one cycle of amplification, which can be conducted using a collection of overlapping single-stranded DNA fragments of related sequence, and different lengths. Fragments can be prepared using a single stranded DNA phage, such as M13 (see, Wang, *Biochemistry* 36:9486-9492 (1997)). Each fragment can hybridize to and prime polynucleotide chain extension of a second fragment from the collection, thus forming sequence-recombined polynucleotides. In a further variation, ssDNA fragments of variable length can be generated from a single primer by Pfu, Taq, Vent, Deep Vent, UITma DNA polymerase or other DNA polymerases on a first DNA template (see, Cline, *Nucleic Acids Res.* 24:3546-3551 (1996)). The single stranded DNA fragments are used as primers for a second, Kunkel-type template, consisting of a uracil-

containing circular ssDNA. This results in multiple substitutions of the first template into the second. *See, Levichkin, Mol. Biology 29:572-577 (1995); Jung, Gene 121:17-24 (1992).*

In some embodiments of the invention, shuffled nucleic acids obtained by use of the recursive recombination methods of the invention, are put into a cell and/or organism for screening. Shuffled monooxygenase genes can be introduced into, for example, bacterial cells, yeast cells, fungal cells vertebrate cells, invertebrate cells or plant cells for initial screening. *Bacillus* species (such as *B. subtilis* and *E. coli* are two examples of suitable bacterial cells into which one can insert and express shuffled monooxygenase genes which provide for convenient shuttling to other cell types (a variety of vectors for shuttling material between these bacterial cells and eukaryotic cells are available; *see, Sambrook, Ausubel and Berger, all supra*). The shuffled genes can be introduced into bacterial, fungal or yeast cells either by integration into the chromosomal DNA or as plasmids.

Although bacterial and yeast systems are most preferred in the present invention, in one embodiment, shuffled genes can also be introduced into plant cells for production purposes (it will be appreciated that transgenic plants are, increasingly, an important source of industrial enzymes). Thus, a transgene of interest can be modified using the recursive sequence recombination methods of the invention *in vitro* and reinserted into the cell for *in vivo/in situ* selection for the new or improved monooxygenase property, in bacteria, eukaryotic cells, or whole eukaryotic organisms.

5. *In vivo* DNA Shuffling Formats

In some embodiments of the invention, DNA substrate molecules are introduced into cells, wherein the cellular machinery directs their recombination. For example, a library of mutants is constructed and screened or selected for mutants with improved phenotypes by any of the techniques described herein. The DNA substrate molecules encoding the best candidates are recovered by any of the techniques described herein, then fragmented and used to transfect a plant host and screened or selected for improved function. If further improvement is desired, the DNA substrate molecules are recovered from the host cell, such as by PCR, and the process is repeated until a desired level of improvement is obtained. In some embodiments, the fragments are denatured and reannealed prior to transfection, coated with recombination stimulating proteins such as *recA*, or co-transfected with a selectable marker such as *Neo^R* to allow the positive selection

for cells receiving recombined versions of the gene of interest. Methods for *in vivo* shuffling are described in, for example, PCT application WO 98/13487 and WO 97/20078.

The efficiency of *in vivo* shuffling can be enhanced by increasing the copy number of a gene of interest in the host cells. For example, the majority of bacterial cells in stationary phase cultures grown in rich media contain two, four or eight genomes. In minimal medium the cells contain one or two genomes. The number of genomes per bacterial cell thus depends on the growth rate of the cell as it enters stationary phase. This is because rapidly growing cells contain multiple replication forks, resulting in several genomes in the cells after termination. The number of genomes is strain dependent, although all strains tested have more than one chromosome in stationary phase. The number of genomes in stationary phase cells decreases with time. This appears to be due to fragmentation and degradation of entire chromosomes, similar to apoptosis in mammalian cells. This fragmentation of genomes in cells containing multiple genome copies results in massive recombination and mutagenesis. The presence of multiple genome copies in such cells results in a higher frequency of homologous recombination in these cells, both between copies of a gene in different genomes within the cell, and between a genome within the cell and a transfected fragment. The increased frequency of recombination allows one to evolve a gene evolved more quickly to acquire optimized characteristics.

In nature, the existence of multiple genomic copies in a cell type would usually not be advantageous due to the greater nutritional requirements needed to maintain this copy number. However, artificial conditions can be devised to select for high copy number. Modified cells having recombinant genomes are grown in rich media (in which conditions, multicopy number should not be a disadvantage) and exposed to a mutagen, such as ultraviolet or gamma irradiation or a chemical mutagen, *e.g.*, mitomycin, nitrous acid, photoactivated psoralens, alone or in combination, which induces DNA breaks amenable to repair by recombination. These conditions select for cells having multicopy number due to the greater efficiency with which mutations can be excised. Modified cells surviving exposure to mutagen are enriched for cells with multiple genome copies. If desired, selected cells can be individually analyzed for genome copy number (*e.g.*, by quantitative hybridization with appropriate controls). For example, individual cells can be sorted using a cell sorter for those cells containing more DNA, *e.g.*, using DNA specific fluorescent compounds or sorting for increased size using light dispersion. Some or all of the collection

of cells surviving selection are tested for the presence of a gene that is optimized for the desired property.

In one embodiment, phage libraries are made and recombined in mutator strains such as cells with mutant or impaired gene products of *mutS*, *mutT*, *mutH*, *mutL*, *ovrD*, *dcm*, *vsr*, *umuC*, *umuD*, *sbcB*, *recJ*, *etc.* The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. High multiplicity of infection (MOI) libraries are used to infect the cells to increase recombination frequency.

Additional strategies for making phage libraries and or for recombining DNA from donor and recipient cells are set forth in U.S. Pat. No. 5,521,077. Additional recombination strategies for recombining plasmids in yeast are set forth in WO 97 07205.

6. Whole Genome Shuffling

In one embodiment, the selection methods herein are utilized in a "whole genome shuffling" format. An extensive guide to the many forms of whole genome shuffling is found in the pioneering application to the inventors and their co-workers entitled "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination," Attorney Docket No. 018097-020720US filed July 15, 1998 by del Cardayre *et al.* (USSN 09/161,188).

In brief, whole genome shuffling makes no presuppositions at all regarding what nucleic acids may confer a desired property. Instead, entire genomes (*e.g.*, from a genomic library, or isolated from an organism) are shuffled in cells and selection protocols applied to the cells.

The fermentation of microorganisms for the production of natural products is the oldest and most sophisticated application of biocatalysis.

The methods herein allow monooxygenase biocatalysts to be improved at a faster pace than conventional methods. Whole genome shuffling can at least double the rate of strain improvement for microorganisms used in fermentation as compared to traditional methods. This provides for a relative decrease in the cost of fermentation processes. New products can enter the market sooner, producers can increase profits as well as market share, and consumers gain access to more products of higher quality and at lower prices. Further, increased efficiency of production processes translates to less waste production and more frugal use of resources. Whole genome shuffling provides a means of accumulating multiple

useful mutation per cycle and thus eliminate the inherent limitation of current strain improvement programs (SIPs).

DNA shuffling provides recursive mutagenesis, recombination, and selection of DNA sequences. A key difference between DNA shuffling-mediated recombination and natural sexual recombination is that DNA shuffling effects both the pairwise (two parents) and the poolwise (multiple parents) recombination of parent molecules. Natural recombination is more conservative and is limited to pairwise recombination. In nature, pairwise recombination provides stability within a population by preventing large leaps in sequences or genomic structure that can result from poolwise recombination. However, for the purposes of directed evolution, poolwise recombination is appealing since the beneficial mutations of multiple parents can be combined during a single cross to produce a superior offspring. Poolwise recombination is analogous to the crossbreeding of inbred strains in classic strain improvement, except that the crosses occur between many strains at once. In essence, poolwise recombination is a sequence of events that effects the recombination of a population of nucleic acid sequences that results in the generation of new nucleic acids that contains genetic information from more than two of the original nucleic acids.

There are a few general methods for effecting efficient recombination in prokaryotes. Bacteria have no known sexual cycle *per se*, but there are natural mechanisms by which the genomes of these organisms undergo recombination. These mechanisms include natural competence, phage-mediated transduction, and cell-cell conjugation. Bacteria that are naturally competent are capable of efficiently taking up naked DNA from the environment. If homologous, this DNA undergoes recombination with the genome of the cell, resulting in genetic exchange. *Bacillus subtilis*, the primary production organism of the enzyme industry, is known for the efficiency with which it carries out this process.

In generalized transduction, a bacteriophage mediates genetic exchange. A transducing phage will often package headfulls of the host genome. These phage can infect a new host and deliver a fragment of the former host genome which is frequently integrated via homologous recombination. Cells can also transfer DNA between themselves by conjugation. Cells containing the appropriate mating factors transfer episomes as well as entire chromosomes to an appropriate acceptor cell where it can recombine with the acceptor genome. Conjugation resembles sexual recombination for microbes and can be intraspecific, interspecific, and intergeneric. For example, an efficient means of transforming

Streptomyces sp., a genera responsible for producing many commercial antibiotics, is by the conjugal transfer of plasmids from *Escherichia coli*.

For many industrial microorganisms, knowledge of competence, transducing phage, or fertility factors is lacking. Protoplast fusion has been developed as a versatile and general alternative to these natural methods of recombination. Protoplasts are prepared by removing the cell wall by treating cells with lytic enzymes in the presence of osmotic stabilizers. In the presence of a fusogenic agent, such as polyethylene glycol (PEG), protoplasts are induced to fuse and form transient hybrids or "fusants." During this hybrid state, genetic recombination occurs at high frequency allowing the genomes to reassort. The final step is the successful segregation and regeneration of viable cells from the fused protoplasts. Protoplast fusion can be intraspecific, interspecific, and intergeneric and has been applied to both prokaryotes and eukaryotes. In addition, it is possible to fuse more than two cells, thus providing a mechanism for effecting poolwise recombination. While no fertility factors, transducing phages or competency development is needed for protoplast fusion, a method for the formation, fusing, and regeneration of protoplasts is typically optimized for each organism.

Modifications can be made to the method and materials as hereinbefore described without departing from the spirit or scope of the invention as claimed, and the invention can be put to a number of different uses, including:

The use of an integrated system to test monooxygenase in shuffled DNAs, including in an iterative process.

7. Family Shuffling P450s

For identification of homologous genes used in family shuffling strategies, representative alignments of P450 enzymes can be found in the Appendices of the volume CYTOCHROME P450: STRUCTURE, MECHANISM, AND BIOCHEMISTRY, 2nd Addition (ed. by Paul R. Ortiz de Montellano) Plenum Press, New York, 1995) ("Ortiz de Montellano"). An up-to-date list of P450s can be found electronically on the World Wide Web (<http://drnelson.utmem.edu/homepage.html>).

To illustrate the family shuffling approach to improving P450 enzymes, one or more of the more than 1000 members of this superfamily is selected, aligned with similar homologous sequences, and shuffled against these homologous sequences.

For example, the gene for the bovine P450_{sc} enzyme, *CYP11A1*, belongs to a family of closely related P450 genes. DNA family shuffling (Cramer *et al.*, *Nature* 391:288) can be used to create hybrid variants from these genes, variants of which can be screened for enhanced conversion of cholesterol to pregnenolone.

5 The screening is done most easily in yeast, but a bacterial system could also be constructed by co-expressing the accessory electron transport proteins adrenodoxin and adrenodoxin reductase. DNA from clones with improved activity can be shuffled together in subsequent rounds of DNA shuffling and screened for further improvement.

10 Subsequent steps in the biosynthesis of steroids such as cortisone and estradiol are also catalyzed by cytochrome P450 enzymes (*see*, Ortiz de Montellano, chapter 12.) For example, conversion of pregnenolone to cortisol involves four enzymatic steps, three of which are catalyzed by cytochrome P450 enzymes. Each of these enzymes belongs to P450 gene families, which also are amenable to DNA family shuffling.

15 One model P450 system has been developed by Pompon and co-workers (*e.g.*, Duport *et al.*, *Nature Biotechnol.* 16:186; Pompon *et al.*, *Methods Enzymol.* 272:51). In particular, they have developed a yeast strain that produces pregnenolone from galactose, and an additional strain that further converts pregnenolone to progesterone. One of the enzymes expressed in these strains is the bovine P450_{sc}. Optimization of this strain, or of related processes useful for steroid production can be assisted by DNA shuffling of P450_{sc}.
20 Numerous other microbial expression systems for P450-type enzymes are known in the literature.

8. Codon Modification Shuffling

Procedures for codon modification shuffling are described in detail in
25 SHUFFLING OF CODON ALTERED GENES, Phillip A. Patten and Willem P.C. Stemmer, filed September 29, 1998, USSN 60/102362 and in SHUFFLING OF CODON ALTERED GENES, Phillip A. Patten and Willem P.C. Stemmer, filed January 29, 1999, USSN 60/117729. In brief, by synthesizing nucleic acids in which the codons encoding polypeptides are altered, it is possible to access a completely different mutational cloud upon
30 subsequent mutation of the nucleic acid. This increases the sequence diversity of the starting nucleic acids for shuffling protocols, which alters the rate and results of forced evolution procedures. Codon modification procedures can be used to modify any nucleic acid

described herein, *e.g.*, prior to performing DNA shuffling, or codon modification approaches can be used in conjunction with oligonucleotide shuffling procedures as described *supra*.

In these methods, a first nucleic acid sequence encoding a first polypeptide sequence is selected. A plurality of codon altered nucleic acid sequences, each of which
5 encode the first polypeptide, or a modified or related polypeptide, is then selected (*e.g.*, a library of codon altered nucleic acids can be selected in a biological assay which recognizes library components or activities), and the plurality of codon-altered nucleic acid sequences is recombined to produce a target codon altered nucleic acid encoding a second protein. The target codon altered nucleic acid is then screened for a detectable functional or structural
10 property, optionally including comparison to the properties of the first polypeptide and/or related polypeptides. The goal of such screening is to identify a polypeptide that has a structural or functional property equivalent or superior to the first polypeptide or related polypeptide. A nucleic acid encoding such a polypeptide can be used in essentially any procedure desired, including introducing the target codon altered nucleic acid into a cell,
15 vector, virus, attenuated virus (*e.g.*, as a component of a vaccine or immunogenic composition), transgenic organism, or the like.

9. Oligonucleotide and *in silico* shuffling formats

In addition to the formats for shuffling noted above, at least two additional
20 related formats are useful in the practice of the present invention. The first, referred to as "in silico" shuffling utilizes computer algorithms to perform "virtual" shuffling using genetic operators in a computer. As applied to the present invention, gene sequence strings are recombined in a computer system and desirable products are made, *e.g.*, by reassembly PCR of synthetic oligonucleotides. In silico shuffling is described in detail in Selifonov and
25 Stemmer in "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed February 5, 1999, USSN 60/118854. In brief, genetic operators (algorithms which represent given genetic events such as point mutations, recombination of two strands of homologous nucleic acids, *etc.*) are used to model recombinational or mutational events
30 which can occur in one or more nucleic acid, *e.g.*, by aligning nucleic acid sequence strings (using standard alignment software, or by manual inspection and alignment) and predicting recombinational outcomes. The predicted recombinational outcomes are used to produce corresponding molecules, *e.g.*, by oligonucleotide synthesis and reassembly PCR.

The second useful format is referred to as "oligonucleotide mediated shuffling" in which oligonucleotides corresponding to a family of related homologous nucleic acids (*e.g.*, as applied to the present invention, interspecific or allelic variants of a dioxygenase nucleic acid) which are recombined to produce selectable nucleic acids. This format is described in detail in Crameri *et al.* "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri *et al.* "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049. The technique can be used to recombine homologous or even non-homologous nucleic acid sequences.

One advantage of the oligonucleotide-mediated recombination is the ability to recombine homologous nucleic acids with low sequence similarity, or even non-homologous nucleic acids. In these low-homology oligonucleotide shuffling methods, one or more set of fragmented nucleic acids are recombined, *e.g.*, with a set of crossover family diversity oligonucleotides. Each of these crossover oligonucleotides have a plurality of sequence diversity domains corresponding to a plurality of sequence diversity domains from homologous or non-homologous nucleic acids with low sequence similarity. The fragmented oligonucleotides, which are derived by comparison to one or more homologous or non-homologous nucleic acids, can hybridize to one or more region of the crossover oligos, facilitating recombination.

When recombining homologous nucleic acids, sets of overlapping family gene shuffling oligonucleotides (which are derived by comparison of homologous nucleic acids and synthesis of oligonucleotide fragments) are hybridized and elongated (*e.g.*, by reassembly PCR), providing a population of recombined nucleic acids, which can be selected for a desired trait or property. Typically, the set of overlapping family shuffling gene oligonucleotides include a plurality of oligonucleotide member types which have consensus region subsequences derived from a plurality of homologous target nucleic acids.

Typically, family gene shuffling oligonucleotide are provided by aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity. A plurality of family gene shuffling oligonucleotides are synthesized (serially or in parallel) which correspond to at least one region of sequence diversity.

Sets of fragments, or subsets of fragments used in oligonucleotide shuffling approaches can be provided by cleaving one or more homologous nucleic acids (*e.g.*, with a

DNase), or, more commonly, by synthesizing a set of oligonucleotides corresponding to a plurality of regions of at least one nucleic acid (typically oligonucleotides corresponding to a full-length nucleic acid are provided as members of a set of nucleic acid fragments). In the shuffling procedures herein, these cleavage fragments (*e.g.*, fragments of monooxygenases) can be used in conjunction with family gene shuffling oligonucleotides, *e.g.*, in one or more recombination reaction to produce recombinant monooxygenase nucleic acids.

10. Chimeric shuffling templates

In addition to the naturally occurring, mutated and synthetic oligonucleotides discussed above, polynucleotides encoding chimeric polypeptide can be used as substrates for shuffling in any of the above-described shuffling formats. Nucleic acids encoding chimeras prepared by art-recognized are encompassed herein. Art-recognized methods for preparing chimeras are applicable to the methods described herein (*see*, for example, Shimoji *et al.*, *Biochemistry* 37: 8848-8852 (1998)).

Thus, in another embodiment, the invention provides a chimeric monooxygenase polynucleotide shuffling template. Preferred templates are derived from the P-450 superfamily of monooxygenases.

Cytochrome P450 constitutes a super family of over 1000 members. These proteins are grouped based on their heme prosthetic group and alignments. The sequence identity between the various P450 families is quite low, but the protein three dimensional folds are very similar. Hence alignments can easily be made between P450's using multiple sequence alignment tools such as clustal, DIALIGN, FASTA, MEME, and Block Maker. If a number of programs are used, a consensus alignment is evident, especially around critical residues such as the cysteine bound to the heme.

There are four P450 crystal structures known, P450 -cam, -terp, -eryF and-BM-P, and they all show similar architecture. Although all of the known crystal structures are for bacterial P450, when alignments are done to mammalian enzymes, predictions about the active site pockets and residues can be made. Site directed mutation studies based upon this scheme have experimentally verified the importance of the predicted residues in substrate binding (Gotoh, *J. Biol. Chem.* 267:83-90) describes a model of CYP 2C9, based on P450cam, which others have used and verified. For use of the BM-P structure to model/mutate CYP 4A proteins, *see*, *J. Biol. Chem.* Sep 4; 273(36):23055-61 (1998).

In another aspect, the invention provides a method of obtaining a polynucleotide that encodes a recombinant P450 polypeptide comprising a backbone domain and an active site domain. The method involves: (a) recombining at least first and second forms of a nucleic acid that encodes a P450 active site domain, wherein the first and second forms differ from each other in two or more nucleotides to produce a library of recombinant active site domain encoding polynucleotides; and (b) linking the recombinant active site domain-encoding polynucleotide to a backbone-encoding polynucleotide so that the active site-encoding domain and the backbone-encoding domain are in-frame.

In yet another aspect, the invention provides a method of obtaining a polynucleotide that encodes a recombinant P450 polypeptide comprising a backbone domain and an active site domain. The method involves: (a) recombining at least first and second forms of a nucleic acid that encodes a P450 backbone domain, wherein the first and second forms differ from each other in two or more nucleotides to produce a library of recombinant backbone domain encoding polynucleotides; and (b) linking the recombinant backbone domain-encoding polynucleotide to a active site-encoding polynucleotide so that the backbone-encoding domain and the active site-encoding domain are in-frame.

In a still further aspect, the invention provides a method of obtaining a polynucleotide that encodes a recombinant P450 polypeptide comprising a backbone domain and an active site domain. The method involves: (a) recombining at least first and second forms of a nucleic acid that encodes a P450 active site domain, wherein the first and second forms differ from each other in two or more nucleotides to produce a library of recombinant active site domain encoding polynucleotides; (b) recombining at least first and second forms of a nucleic acid that encodes a P450 backbone domain, wherein the first and second forms differ from each other in two or more nucleotides to produce a library of recombinant backbone domain encoding polynucleotides; and (c) linking the recombinant active site domain-encoding polynucleotide to the recombinant backbone-encoding polynucleotide so that the recombinant active site-encoding domain and the recombinant backbone-encoding domain are in-frame.

The linking of the various nucleic acids in each of the above aspects can be accomplished by methods well-known in the art. Moreover, in each of the above aspects, certain embodiments are presently preferred. For example, in a preferred embodiment, the backbone P450 (BM-P in this example) refers to the C-terminus of the protein which contains the proximal cysteine (residue 400) ligand to the prosthetic heme. The N terminus

of the desired P450 isozyme is transferred onto this structure. In a preferred embodiment the junction between the two sequences occurs at an end of the I helix (e.g., residue 282). In another preferred embodiment the junction between the two proteins occurs in the G-H loop (residues 227-232 preferably). In another preferred embodiment solely the F and G helices
5 (residues 171-226) are transferred into the backbone P450 with the remaining sequence being from the backbone P450.

Using the above methods, chimeric monooxygenases having optimized activities can be obtained. The activities that are optimized include any of the activities towards any of the substrates described herein.

10 Generating a focused P450 library of chimeras, steroid hydroxylases for example, typically begins with an investigation of the literature, especially the drug metabolism area, for isozymes known to catalyze the desired chemistry. Once identified, these isozymes are aligned, using the relevant programs, to one of the P450's with a known x-ray structure (P450 -cam, -terp, -eryF and -BM-P), preferably BM-P. Once the alignment
15 is achieved, the putative active site regions are generated and isolated for further study.

Inspection of the published structures for P450's (see, for example *P.N.A.S.* 96: 1863-1868 (1999); *Nature Struct. Biol.* 4: 140-146 (1997)) and structure function studies (see, for example, *Drug Metab. Dispos.* 26: 1223-1231 (1998), for a review) and are used to highlight the sites at which chimeras are preferably constructed. For the purpose of clarity,
20 all residue numbers refer to an exemplary sequence, CYP 102 P450 BM-P. This focus is not intended to limit the invention as it is apparent that it is the positions in the structural motif of the protein that are relevant not the absolute residue number. The positions of the structural motifs may be determined by methods including crystal structure determination, sequence alignment and homology modeling. Indeed a small extension of the sequence
25 beyond the chosen region may be transferred into the chimera.

The method provides a series of chimeric nucleic acids which include sequences, chosen as described above, from the P450 isozymes known to catalyze the desired chemistry and the remainder of a soluble bacterial P450, preferably one of the structurally defined P450s, most preferably P450BM-P, most preferably still an already
30 improved chimeric monooxygenase nucleic acid. These chimeric nucleic acids can be used as substrates for shuffling in any of the above-described shuffling formats.

In one embodiment the entire polynucleotide is improved by shuffling. In a preferred embodiment, the heme domain of the P450 component of the chimera is shuffled.

In another preferred embodiment the active site region of the P450 isozymes is shuffled. In yet another preferred format the active site sequences described above are shuffled before chimera formation. In this format the improved nucleic acids are cloned into the P450 backbone to create a library of improved monooxygenases

5 In another preferred format, one or more of the desired P450 isozyme active sites are not transformed into a chimeric nucleic acid. The diversity encoded by these sequences are captured by the inclusion of oligonucleotides encoding the sequence of interest as described in the above-described shuffling format.

One advantage of this process is that the formation of chimeric P450
10 nucleotides allows the production of polypeptide encoding any P450 activity in the same system. Thus the creation of an improved nucleic acid with one activity may start from a previously improved chimeric nucleic acid encoding a different activity. This recursive synergy leads to rapid improvement of the monooxygenase nucleic acid for any and all of the desired properties.

15 Another advantage of this process is the improvement in stability and ease of expression of polypeptides with the activity of a eukaryotic, membrane associated, P450 as a soluble bacterial protein. This leads to significant improvement in the expression level, stability, and ease of handling of any polypeptide encoded by the improved nucleic acid.

A third advantage of this process is the ability to create improved nucleic
20 acids for a particular activity without isolation of the nucleic acid encoding that activity. Each chimeric nucleic acid will be expressed and screened in substantially similar fashion for any of the reactions described herein.

Thus any reaction described in the literature of biotransformation and drug metabolism and known to those skilled in the art, such as those described herein, encoded by
25 a P450 nucleic acid can be performed by a chimeric nucleic acid of the type described.

B. Reactions of Improved Monooxygenases

In another aspect, the invention provides a method for obtaining a polynucleotide encoding an improved polypeptide acting on a substrate comprising a target
30 group selected from an olefin, a terminal methyl group, a methylene group, an aryl group and combinations thereof. The improved polypeptide exhibits one or more improved properties compared to a naturally occurring polypeptide acting on said substrate. The method includes: (a) creating a library of recombinant polynucleotides encoding a

monooxygenase polypeptide acting on said substrate; and (b) screening said library to identify a recombinant polynucleotide encoding an improved polypeptide that exhibits one or more improved properties compared to a naturally occurring monooxygenase polypeptide.

In a preferred embodiment, the library of recombinant polynucleotides is created by recombining at least a first form and a second form of a nucleic acid. At least one of these forms encodes the naturally occurring polypeptide or a fragment thereof. Preferably, the first form and said second form differ from each other in two or more nucleotides. In a further preferred embodiment, the first and second forms of the nucleic acid are homologous.

In addition to the methods described above for producing the encoding polynucleotides, the present invention also provides the polypeptides encoded by these polynucleotides and methods using these peptides for synthesizing valuable organic compounds. Some of these polypeptides and methods of using them are set forth below.

It is noted that the basic chemistry described below with reference to monooxygenases is known. In addition to Ortiz de Montellano, *supra*, a general guide to the various chemistries involved is found in Stryer (1988) BIOCHEMISTRY, third edition (or later editions) Freeman and Co., New York, NY; Pine *et al.* ORGANIC CHEMISTRY, FOURTH EDITION (1980) McGraw-Hill, Inc. (USA) (or later editions); March, ADVANCED ORGANIC CHEMISTRY REACTIONS, MECHANISMS and Structure, 4th ed, J. Wiley and Sons (New York, NY, 1992) (or later editions); Greene, *et al.*, PROTECTIVE GROUPS IN ORGANIC CHEMISTRY, 2nd Ed., John Wiley & Sons, New York, NY, 1991 (or later editions); Lide (ed) THE CRC HANDBOOK OF CHEMISTRY AND PHYSICS 75TH EDITION (1995)(or later editions); and in the references cited in the foregoing. Furthermore, an extensive guide to many chemical and industrial processes applicable to the present invention is found in the KIRK-OTHMER ENCYCLOPEDIA OF CHEMICAL TECHNOLOGY (third edition and fourth edition, through year 1998), Martin Grayson, Executive Editor, Wiley-Interscience, John Wiley and Sons, NY, and in the references cited therein ("Kirk-Othmer").

The following chemistries illustrate those generally accessible through the heme-dependent P450 monooxygenase/oxidase superfamily. Certain useful reaction types are set forth in Fig 1.

Family shuffling approaches apply to enhancing performance of monooxygenase polypeptides useful in each of the following classes of industrial chemical transformation. Other monooxygenase enzyme classes are also useful in practicing the

present invention. Moreover, other polypeptides accessible through the present invention, and method of using these polypeptides will be apparent to those of skill in the art.

1. Oxidation of π -bonds to epoxides

5 Among the most high-value classes of commodity chemical transformations is the catalytic epoxidation of terminal olefins to corresponding epoxides. Indeed, ethylene oxide, propylene oxide, epichlorohydrin, glycidol, butylene oxide and bis-A-diglycidyl ethers and their immediate downstream derivatives account for a significant fraction of the entire \$350 B/yr global chemical industry. Typically, prior art P450 activities are limited by
10 low turnover number, low affinity, low stability under the conditions of interest and/or enzyme inactivation by alkylation or free-radical-dependent mechanisms. Moreover, such chemistry is often associated with rapid inactivation of the heme-dependent enzyme. Family shuffling approaches to enzyme improvement are used to markedly reduce the sensitivity of the monooxygenases to this mode of inactivation.

15 In a preferred embodiment, the present invention provides an improved polypeptide that is capable of converting an olefin into an epoxide. Moreover, there is provided a method for converting an olefin to an epoxide. The method includes contacting the olefin substrate with the polypeptide. In a still further preferred embodiment, the substrate is contacted with an organism that expresses the polypeptide.

20 In another preferred embodiment, the polypeptides are those encoded by monooxygenase genes that can be recruited and optimized by DNA shuffling. A range of monooxygenases known in the art provide appropriate starting points for determining a polypeptide useful in this aspect of the invention. One useful class of monooxygenases is exemplified by the heme-dependent eukaryotic and bacterial cytochrome P-450

25 Heme-containing enzymes of the P450 family exhibit a wide array of catalytic activities of interest in the context of metabolizing xenobiotics and environmental and biochemical waste products. Of the diverse chemistries catalyzed by this class of enzymes, a number are of industrial chemical interest.

30 As an enzyme class, the P450 family exhibits notable activities toward many classes of compounds. For example, in the presence of oxygen and an intact redox recycle system, P450s exhibit monooxygenase activity. Addition of hydrogen peroxide or other peroxides, however, can be used to circumvent the NAD(P)H requirement (*i.e.* allowing for peroxidase activity) toward many of the same substrates.

In a further preferred embodiment, polypeptides based on, or analogous to, non-heme-dependent monooxygenases are used to effect epoxidation of olefins. Such monooxygenases include, but are not limited to, non-heme monooxygenases involved in the bacterial degradation of styrene by bacteria (as exemplified by the genes and enzymes described by Marconi *et al.*, *Appl. Environ. Microbiol.* **62**(1):121-127 (1996); Beltrametti *et al.*, *Appl. Environ. Microbiol.* **63**(6):2232-2239 (1997); O'Connor *et al.*, *Appl. Environ. Microbiol.* **63**(11):4287-4291 (1997); Velasco *et al.*, *J. Bacteriol.* **180**(5):1063-1071 (1998); Itoh *et al.*, *Biosci. Biotechnol. Biochem.* **60**(11):1826-1830 (1996)), or in the degradation of methyl-substituted aromatic compounds such as toluene, xylenes, *p*-cymene (exemplified by xylene monooxygenase, Wubbolts *et al.*, *Enzyme Microb. Technol.* **16**(7):608-615 (1994)).

The following is a non-limiting list of exemplary monooxygenase genes which can be recruited and optimized by DNA shuffling for the purpose of epoxidizing olefins:

[AF031161] styrene monooxygenase (epoxide-forming) of *Pseudomonas* sp. VLB120, stdA, stdB; [PFSTYABCD] styrene monooxygenase of *P. fluorescens* (styA, styB); [PSSTYCATA] styrene monooxygenase of *Pseudomonas* sp.; [PSEXYLMA, AF019635, D63341, E02361] xylene/toluene monooxygenase of *Pseudomonas putida* TOL plasmid (xyl M, xylA); [PPU24215] *p*-cymene monooxygenase of *P. putida*; [PSETBMAF] toluene/benzene-2-monooxygenase (tbmA-tmmF) of *Pseudomonas* sp.; [PPU04052] toluene-3-monooxygenase of *Pseudomonas pickettii* PKO1; [AF001356] toluene-3-monooxygenase of *Burkholderia cepacia*; and [AF043544] nitrotoluene monooxygenase of *Pseudomonas* sp. TW3, NtnMA (ntnM, ntnA).

A variety of strains known to contain monooxygenases capable of epoxide formation are known. For example, *Pseudomonas aeruginosa* is known to have a monooxygenase capable of epoxidizing 1-octene to 1,2-epoxyoctane. The most comprehensive studies on bacterial alkene epoxidation have been done on *Pseudomonas oleovorans*. Work on *P. oleovorans* by May and coworkers (*J. Biol. Chem.* **248**:1725-1730, 1973) shows that the monooxygenase contained in the cells is capable of epoxidizing octene to 1,2-epoxy-octane in 70% enantiomeric purity. In addition, this enzyme is capable of converting 1,7-octadiene to the diepoxide (May *et al.*, *J. Am. Chem. Soc.* **98**:7856-7858) and 1,5-hexadiene and 1,11-dodecadiene to epoxides. However, smaller alkenes are often

converted to alcohols. Cells grown up overnight under standard conditions can be used intact or as lysates—and, in both cases, have been observed to give yields of ~ 1 g/L. Increasing the rate of accumulation of the reactive epoxide is clearly one of the preferred objectives of gene shuffling as set forth herein.

5 This enzyme system is also capable of mediating hydroxylation of longer chain alkanes (octanes, *etc.*) and fatty acids. The enzyme has been cloned and sequenced and is included of three protein components: rubredoxin (mw 19,000), NADH-rubredoxin reductase, and the hydroxylase (a non-heme iron protein). Whereas there are scenarios (such as when overall stability of the system is an issue) in which shuffling of the genes for all
10 three protein components is preferred, when the primary improvement is related to the kinetics, affinity or inhibition profile of the monooxygenase, the preferred shuffling strategy will be to shuffle homologs of the hydroxylase (epoxygenase) component.

 Microorganisms having MO enzyme activities with similar properties include the genera *Rhodococcus*, *Mycobacterium*, *Nocardia* (*Nocardia corollina* B-276) and
15 *Pseudomonas Corynebacterium equi* (IFO 3730), which can be grown on n-octane and which exhibit the capacity to oxidize 1-hexene to optically pure (R)-(+)-epoxide. This strain also assimilates other terminal olefins and converts them to epoxides. Yields decrease to <1% with carbon chains of >14. Increasing the activity of the enzyme toward longer chain length alkenes is a target for evolving additional catalysts for chirally selective
20 epoxidations. Such monomers have high value as pharmaceutical and agricultural intermediates.

 Experiments with *Pseudomonas putida*, *Nocardia corallina* B-276 and *Bacillus megaterium*, suggest that the monooxygenase activity of these organisms derives from a soluble P450-dependent system. All of these strains are available from ATCC and
25 serve as exemplary sources for the genes which can be isolated by hybridization and gene amplification methods.

Mycobacterium sp (E20) and *Mycobacterium sp.* (Py 1) show activity even toward short-chain, gaseous olefins such as ethylene. In the case of both ethylene and propylene, the epoxide products are formed almost exclusively. Catalyst performance
30 experiments are performed in a gas-solid reactor to prevent accumulation of toxic ethylene oxide in the immediate vicinity of the biocatalyst. An experimental set-up which allows for automatic gas chromatography analysis of circulation gas in a batch reactor system and allows for online monitoring of the microbial (or enzymatic) oxidation of gaseous alkenes

(ethylene, propylene and butylene). Optimization of the process is achieved by studying the influence of various organic solvents and physical conditions on retention of immobilized cell/enzyme activity.

High activity retention is favored by low polarity, high molecular weight solvents; although this is also selectable following DNA shuffling as well. Using chiral gas chromatography, wild type (wt) strains and strains containing candidate evolved polypeptides are screened with respect to the stereospecificity of the epoxidation of propene, 1-butene and 3-chloro-1-propene. Results show that a wide range of chiral selectivity or nonselectivity emerge from a typical series of family shuffling and screening experiments. Novel polypeptides, favoring the S, rather than the R stereoisomer can also be shuffled and selected. Inactivation of the alkene epoxidation system by the produced epoxide has been one of the key historical limitations of the system. Again, gene and family shuffling combined with appropriate selection methods and screens are used to identify polypeptides with improved stability in the presence of epoxide products.

A number of other methane-grown methylotrophic bacteria (*Methylosinus trichosporium*, *Methylobacterium capsulatus* and *Methylobacterium organophilum*) have all been shown to contain a methane monooxygenase (MMO) system analogous to the well-characterized *Pseudomonas oleovorans* system. Again, standard hybridization and gene amplification methods provide a straightforward approach to isolate those genes which are not yet reported in the literature. Sequences of MMOs from some of these organisms are known and can be obtained from the public sequence Databases such as Genbank, Entrez®, and others.

Moreover, one species of *Rhodococcus rhodochrous* has been shown to be capable of oxidizing propane and propene to epoxide and hydroxylated products without inhibition by the products. The unique monooxygenase from this organism provides an important material to incorporate in family shuffling formats to expand activity of shuffled nucleic acids.

2. Hydroxylation of organic substrates

In another embodiment, the present invention provides a monooxygenase polypeptide capable of hydroxylating organic substrates. In an exemplary embodiment, the polypeptide oxidizes a methyl or a methylene group. In a preferred embodiment, the polypeptide oxidizes a terminal methyl group to a hydroxymethyl group. In yet another

preferred embodiment, the invention provides an improved monooxygenase polypeptide that acts on a methylene group to form a secondary alcohol. Preferred organic substrates include a target group selected from arylmethyl, substituted arylmethyl, arylmethylene, substituted arylmethylene, heteroaryl methyl, substituted heteroaryl methyl, alkyl-terminal methyl, fatty acid, terpenes and combinations thereof. The improved polypeptide is prepared using the methods of the invention and exhibits one or more improved properties compared to a naturally occurring polypeptide.

In addition to the polypeptide, there is provided a method for converting a terminal methyl or internal methylene into the corresponding alkyl hydroxy group. The method includes contacting the substrate with the polypeptide. In a still further preferred embodiment, the substrate is contacted with an organism that expresses the polypeptide.

P450s mediate the conversion of many of the molecular species listed above, including oxidation of toluene to form benzyl alcohol and oxidation of 2-phenyl-propane to 2-phenyl-1-propanol. Monooxygenase enzymes from *Pseudomonas gladioli*, *Aspergillus niger* and other species are known to oxidize monoterpenes as well as higher terpenes. Conversion of monoterpenes to terminal unsaturated alcohols (without disruption of alkene functionalities) is a remarkable aspect of monooxygenase mediated conversions (see, ENZYME CATALYSIS IN ORGANIC SYNTHESIS, VOL. II, Chapter B.6.1.4 (ed. By K. Drauz and H. Waldmann, VCH Publishers, Inc., 1995). The powerful monooxygenase system of *Pseudomonas oleovorans* is also known to transform linear and branched-chain alkanes to alcohols, aldehydes, acids and hydroxy acids.

Members of the P450 superfamily typically favor formation of primary alcohols. An example of a P450-mediated hydroxylation of interest is the ω and ω -1 hydroxylation of fatty acids, such as lauric acid. P450s such as CYP2B4, CYP2B1 and related sequences demonstrate this activity toward a number of hydrocarbon substrates. Shuffling members of this subfamily leads to polypeptides with altered specificity and enhanced stability.

Many polypeptides capable of arylmethyl group oxidation are well known in the art. For example, the introduction of oxygen into methyl groups and methylene groups is mediated by non-heme multicomponent monooxygenases of toluene, xylenes and *p*-cymene.

While much of the discussion above focuses on constructing polypeptides and pathways for oxidation of arylmethyl compounds, this discussion is also directly applicable to polypeptides and pathways for oxidizing terminal methyl and internal methylene groups

of both alkyl and aryl-substituted alkyl groups. In a preferred embodiment, the substrate is an aryl-substituted alkyl group (*see*, Fig. 2).

This step is accomplished by recruiting one or more genes encoding an appropriate monooxygenase activity. In a preferred embodiment, this is accomplished by shuffling and expressing a suitable cytochrome P450 type enzyme system. The enzymes of this class are ubiquitous in nature, and they can be found in a variety of organisms. For example, n-propylbenzene is known to undergo α -oxidation in strains of *Pseudomonas desmolytica* S449B1 and *Pseudomonas convexa* S107B1 (Jigami *et al.*, *Appl. Environ. Microbiol.* 1979 38(5):783-788).

Similarly, alkane monooxygenases of bacterial origin, or cytochromes P450 for camphor oxidation, whether wild-type or mutant, can be recruited for the purpose of introducing the oxygen into the terminal methyl group of alkylaryl compounds, wherein the alkyl group is generally other than a methyl group (Lee *et al.*, *Biochem. Biophys. Res. Commun.*; 218(1):17-21 (1996); van Beilen *et al.*, *Mol. Microbiol.*; 6(21):3121-3136 (1992); Kok *et al.*, *J. Biol. Chem.* 264(10):5435-5441 (1989); Kok *et al.*, *J. Biol. Chem.* 264(10):5442-5451 (1989); Loida and Sligar, *Protein Eng.* 6(2):207-212 (1993)).

Furthermore, the mammalian metabolic pathways for these and structurally related alkylaromatic hydrocarbons indicate a cytochrome P450 dependent chiral oxidation of the terminal methyl group and subsequent oxidation to corresponding 2-arylpropanoic or 2-arylacetic acids, indicating that these P450s are excellent shuffling substrates (Matsumoto *et al.*, *Chem. Pharm. bull. (Tokyo)* 40(7):1721-1726 (1992); Matsumoto *et al.*, *Biol. Pharm. Bull.* 17(11):1441-1445 (Nov 1994); Matsumoto *et al.*, *Chem. Pharm. Bull. (Tokyo)* 43(2):216-222 (1995); Ishida and Matsumoto, *Xenobiotica* 22(11):1291-1298 (1992)).

Examples of monooxygenase genes suitable for use in the construction of strains for oxidation of the methylarenes include:

[PSEXYLMA, AF019635, D63341, E02361] xylene/toluene monooxygenase of *Pseudomonas putida* TOL plasmid (xyl M, xylA); [PPU24215] *p*-cymene monooxygenase of *P. putida*; [AF043544] nitrotoluene monooxygenase of *Pseudomonas* sp. TW3, NtnMA (ntnM, ntnA); [SMU40233 and SMU40234] alkane monooxygenase of *Stenotrophomonas maltophilia*; [POOCT] alkane monooxygenase of *Pseudomonas oleovorans* TF4-1L (+OCT) plasmid, *alk* genes; and camphor 5-monohydroxylase of *P. putida* (CAM plasmid)

Alternatively, for the purpose of using of non-heme-dependent oxidation of the arylalkyl compounds, useful monooxygenases are exemplified by a variety of non-heme monooxygenases involved in the bacterial degradation of styrene by bacteria (as exemplified by the corresponding genes and enzymes described by Marconi, *et al.*, *Appl. Environ. Microbiol.* **62**(1):121-127 (1996); Beltrametti, *et al.*, *Appl. Environ. Microbiol.* **63**(6):2232-2239 (1997); O'Connor, *et al.*, *Appl. Environ. Microbiol.* **63**(11):4287-4291 (1997); Velasco, *et al.*, *J. Bacteriol.* **180**(5):1063-1071 (1998); Itoh, *et al.*, *Biosci. i Biotechnol. Biochem.* **60**(11):1826-1830 (1996)); or in the degradation of methyl-substituted aromatic compounds such as toluene, xylenes, *p*-cymene (exemplified by xylene monooxygenase, Wubbolts, *et al.*, *Enzyme. Microb. Technol.* **16**(7):608-615 (1994)).

Exemplary non-heme monooxygenases useful in practicing the present invention include:

[AF031161] styrene monooxygenase (epoxide-forming) of *Pseudomonas* sp. VLB120, stdA, stdB, [PFSTYABCD] styrene monooxygenase (epoxide-forming) of *P. fluorescens* (styA, styB); [PSSTYCATA] styrene monooxygenase (epoxide-forming) of *Pseudomonas* sp; [PSEXYLMA, AF019635, D63341, E02361] xylene/toluene monooxygenase of *Pseudomonas putida* TOL plasmid (xyl M, xylA); [PPU24215] *p*-cymene monooxygenase of *P. putida*; [PSETBMAF] toluene/benzene-2-monooxygenase (tbmA-tmmF) of *Pseudomonas* sp.; [PPU04052] toluene-3-monooxygenase of *Pseudomonas pickettii* PKO1; [AF001356]; toluene-3-monooxygenase of *Burkholderia cepacia*; [AF043544] nitrotoluene monooxygenase, of *Pseudomonas* sp. TW3, NtnMA (ntnM, ntnA).

3. Aromatic hydroxylation

Hydroxylated aromatic compounds are an important group of industrial chemicals. Carboxylic acids, esters and lactones of hydroxylated aromatic compounds are of particular value and interest. Thus, in another preferred embodiment, the invention provides an improved monooxygenase polypeptide that can oxidize an aryl compound to a hydroxyaryl compound (Fig. 1). Additionally, there is provided a method utilizing an improved monooxygenase polypeptide to effect the transformation of an aryl group to a heteroaryl group. The method includes contacting a substrate comprising an aryl group with

the polypeptide. In yet another preferred embodiment, the substrate is contacted with an organism that expresses the polypeptide.

Presently preferred substrates include, for example, aryl groups, substituted aryl groups, heteroaryl groups and substituted heteroaryl groups. Compounds representative of these generic groups include industrially significant substrates such as biphenyl, benz-[a]-pyrene, aniline, toluene, naphthalene, cumene, haloaromatics and phenanthrene.

Many monohydroxy aromatic compounds can be generated by using heme- and/or non-heme-containing type monooxygenases. To be useful in the biotransformation pathway, preferred polypeptides will have a sufficiently high turnover rate and they will not be readily deactivated in the presence of the substrates, intermediates or products of the oxidation reaction. This characteristic is an ideal candidate for improvement by the shuffling process disclosed herein.

This class of reactions includes, for example, the modification of such industrially significant substrates as benzene, biphenyl, benz-[a]-pyrene, aniline, toluene, naphthalene, cumene, haloaromatics and phenanthrene are all of considerable industrial chemical importance and are all carried out by members of the P450 superfamily.

4. *S-dealkylation of alkylsulfur compounds*

S-Dealkylation of reduced thio-organics, such as oxidation of parathion can be mediated by the use of improved monooxygenases. Sulfoxidation of numerous organosulfur compounds is also observed and can be enhanced by shuffling monooxygenases. Thus, in another preferred embodiment, the invention provides an improved monooxygenase polypeptide that can oxidize a penicillin G to penicillin G S-oxide, a key intermediate in the synthesis of cephalosporins.

5. *O-Dealkylation of alkyl ethers*

Whereas S and N-alkyl groups are oxidized by monooxygenases to the corresponding oxides, the electronegativity of oxygen dictates a different mechanistic pathway, namely rearrangement of the O-alkyl bond. Synthetic pathways utilizing this reaction motif can be improved by shuffling monooxygenases.

6. *Oxidation of aryloxy phenols*

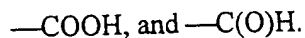
Monooxygenase mediated reactions such as the conversion of p(p-nitrophenoxy)phenol to quinone can be enhanced by shuffling monooxygenases.

7. *Dehydrogenation*

In some cases, the monooxygenase polypeptides of the invention operate as dehydrogenases rather than as oxygenases or peroxidases. For example, conversion of saturated hydrocarbons to unsaturated, conversion of alcohols to aldehydes, carboxylic acids and ketones, conversion of aldehydes to carboxylic acids and the desaturation of nitrogen compounds has been observed. A classic example of this is the conversion of dihydronaphthalene to naphthalene. Conversion of valproic acid to 2-n-propyl-pentenoic acid also illustrates this chemistry as does conversion of lindane (1,2,3,4,5,6-hexachlorocyclohexane) to hexachlorocyclohexene. Numerous other examples of this classic P450 chemical transformation exist, such as conversion of acetaldehyde or propionaldehyde to acetic and propionic acid, respectively. The CYP2C29 enzyme, for example, converts aliphatic alpha-beta unsaturated aldehydes (and anthraldehyde) to the corresponding acids. Shuffling of these and related P450s provides improved properties, such as enhanced activity, specificity and/or P450 stability.

Moreover, P450-based dehydrogenation chemistry also plays an important role in the biosynthesis of various steroids, and is, therefore, of considerable commercial interest in synthesizing steroid-based pharmaceuticals such as cortisol and other steroidal anti-inflammatory agents.

Thus, in another embodiment, the present invention provides a method for obtaining a nucleic acid encoding an improved monooxygenase polypeptide having dehydrogenase activity. In a preferred embodiment, the improved polypeptide acts on a substrate to dehydrogenate a hydroxyalkyl group to a member selected from:



Preferred substrates include members selected from the group of arylmethyl, substituted arylmethyl, heteroarylmethyl, substituted heteroarylmethyl, alkyl-terminal methyl, substituted alkyl-terminal methyl, and the like, as well as combinations thereof.

The improved polypeptide of the invention exhibits one or more improved properties compared to a naturally occurring polypeptide. Producing the polypeptide by the method of the invention involves creating a library of recombinant polynucleotides encoding a polypeptide acting on the substrate; and screening the library to identify a recombinant polynucleotide encoding the improved polypeptide.

Moreover, there is provided a dehydrogenase polypeptide prepared by the method of the invention. A method for utilizing this polypeptide to oxidize a hydroxyalkyl group using the polypeptide is also provided. The method involves contacting a substrate having a hydroxyalkyl group with a polypeptide of the invention, more preferably with an organism expressing a polypeptide of the invention.

8. Decarbonylation

Examples of this important chemistry include conversion of cyclohexanecarboxaldehyde to cyclohexane and formic acid. Conversion of isobutyraldehyde, trimethylacetaldehyde, isovaleraldehyde, 2-methyl-butyraldehyde, citronellal and 2-phenyl-propionaldehyde to their corresponding decarbonylated products are also observed. This chemistry is not observed with unbranched aldehydes such as propionaldehyde and valeraldehyde. This is an important class of catalytic chemistry not easily duplicated abiotically. CYP2B4 is a preferred target for shuffling to improve the native activity of this P450. Shuffling of this family of P450 MOs results in polypeptides with activity toward unbranched aldehydes such as adipaldehyde, valeraldehyde and/or propionaldehyde.

10. Oxidative dehalogenation of haloaromatics and halohydrocarbons

Exemplary substrates for these reaction include, polychlorobenzenes, trichloroethylene, di and trichloro propane, 1,2 dichloroethane and 1,2 1,3 and 1,4 dihydroketones.

11. Baeyer-Villiger monooxygenation

This reaction involves the oxidation of aromatic, open-chain and cyclic ketones to esters and lactones.

12. Exemplary embodiments utilizing monooxygenases

a. Cyclosporin

Cyclosporin A is a nonribosomal peptide drug with antifungal and immunosuppressive properties that is widely used as an immunosuppressant after transplant surgery. There currently exist at least 25 cyclosporin derivatives with various properties, and there is a great demand for new cyclosporin molecules. The creation of new derivatives, however, has been hampered by the difficult synthetic chemistry of these large natural product molecules (MW ~1200). Therefore, a means of overcoming this limitation of traditional chemistry is of great value.

Cytochrome P450 and other monooxygenase enzymes provide an alternative method of making modified cyclosporins. The P450 3A subfamily contains members with various activities on cyclosporin A; for example, the 3A5 enzyme can hydroxylate the amino acid at position 1, and 3A4 can hydroxylate amino acids 1 and 9 as well as demethylate position 4 (Aoyama *et al.*, JBC 264:10388). Other activities exist among the large 3A subfamily, consisting of at least 30 members (*see*, <http://drnelson.utm.edu/homepage.html>).

Alignment of 14 of these 3A genes shows homologies of 67-99%. Such diversity is ideal for shuffling, and provides a means of creating additional genetic diversity in the form of P450 libraries, with concomitant enzymatic diversity. Initial screening for new or improved activities can be done in bacteria, as the human 3A4 enzyme and its accessory reductase are functional in *E. coli* (Parikh *et al.*, *Nature Biotechnol.* 15:784). Activity of clones in libraries can be measured by high throughput mass spectroscopy detection of product molecules, for example. DNA from clones with improved activity can be isolated and shuffled to recombine beneficial mutations, followed by screening for even better activity.

b. Pravastatin

Pravastatin is a steroid drug which lowers serum cholesterol by competitive inhibition of the cholesterol biosynthetic enzyme HMG-CoA reductase. Pravastatin (marketed as Pravachol™ by Bristol-Myers Squibb) is produced by a two-step fermentation (Serizawa *et al.* IN BIOTECHNOLOGY OF ANTIBIOTICS 2ND EDITION, W.R. Stfohl (ed.) (1997) New York: Mascel-Dekker, pp. 777-805): production of the precursor mevastatin by

Penicillium citrinum, and then hydroxylation of mevastatin to pravastatin by a cytochrome P450 enzyme in *Streptomyces carbophilus*.

This invention provides a method to make the second step of this synthesis more efficient by increasing the ability of the *S. carbophilus* P450 to hydroxylate mevastatin.

- 5 The value of this improvement is in decreasing the cost of drug synthesis; much work has already gone into optimizing culture conditions (Serizawa *et al.*, 1997), an indication that it is an expensive process.

The P450 that converts mevastatin to pravastatin has been characterized in some detail (Watanabe *et al.*, *Gene* 163:81-85. (1995)). The gene *cytP-450_{sca-2}* has been
 10 cloned and shows homology to other bacterial P450 genes, including 78% identity with the *S. griseolus* gene *suaC*, whose product is involved in herbicide detoxification (Omer *et al.*, *Nature* 288-291 (1998)), and over 50% identity with several other P450 genes (see below). *CytP-450_{sca-2}* is functional when overexpressed in the laboratory strain *S. lividans*.

15 Table 1. DNA homology between selected cytochrome P450 genes.

CYP105A1 suaC	CYP105D1 soyC	CYP105B1 subC	CYP105A2	Sca2	
--	58%	51	56	78	105A1
	--	51	48	57	105D1
		--	56	52	105B1
			--	53	105A2
				--	Sca2

- Improvement of the ability of *CytP-450_{sca-2}* to convert mevastatin to pravastatin can be accomplished by DNA shuffling. The known sequences provide an ideal platform for the family shuffling technique, wherein related, functional genes are shuffled together to
 20 create the initial library for screening/selection. Some of these genes can be obtained directly from the microbe in which they were identified (*e.g.*, CYP105A1 and CYP105B1 from *S. griseolus* strain ATCC11796, see Omer *et al.*, 1990). Others genes such as *CytP-450_{sca-2}* can be assembled from synthetic oligonucleotides. The initial family shuffling can be done as described (Crameri *et al.*, 1998). The initial screen for improved clones can be
 25 done in a surrogate host, such as *E. coli* or *S. lividans*; cells can be cultured in mevastatin

(or the related compound ML-236B•Na; see Watanabe *et al.*, 1995, *above*) and the production of pravastatin detected by high throughput techniques, probably mass spectroscopy. The hydroxy group will easily differentiate the product from the substrate. The genes can be rescued from the best clones and shuffled together in subsequent cycles.

- 5 The final test would be in an environment resembling actual fermentation conditions as much as possible.

c. Herbicide Resistance and Bioremediation

- One set of P450 gene products with activity against herbicides consists of
- 10 SuaC (CYP105A1) and SubC (CYP105B1) from *Streptomyces griseolus* (Omer *et al.*, *J. Bacteriol.* 172:3335) and related genes from other bacteria. These enzymes are active against sulfonylurea herbicides such as chlorimuron ethyl, chlorsulfuron, and sulfomethuron methyl (Harder *et al.*, *Mol. Gen. Genet.* 227:238). Related bacterial P450 genes have been identified, with DNA sequence homologies of 48-78% (*see*, Table 2 below). Because these
- 15 genes are of bacterial origin, they are best suited to bioremediation uses but may also be useful for creating herbicide-resistant plants.

- Another set of P450 genes can be isolated from plants with herbicide detoxification activities. Such activities are known to be due to plant cytochrome P450s (Lau and O'Keefe, *Methods Enzymol.* 272:235). It is possible to identify the genes, or at
- 20 least portions of them, by using PCR primers targeted to conserved regions of P450s (Holton and Lester, *Methods Enzymol.* 272:275) which are responsible for this activity.

- DNA family shuffling (Cramer *et al.*, *Nature* 391:288) can be used to create hybrid variants from these genes, variants which can be screened for increased herbicide metabolism (detoxification). One way to screen for such activity in large numbers of
- 25 samples is by measuring loss of fluorescence due to metabolism of the fluorescent sulfonylurea W5822 (DuPont) (*see*, Harder *et al.*, *Mol. Gen. Genet.* 227:238). Other suitable screening systems employ mass spectroscopy, HPLC and other well-known analytical methods. Improved clones can be shuffled together in the next cycle of DNA shuffling for further improvement. The best genes can then be transferred to plants and tested for
- 30 conferral of herbicide resistance; further optimization may be necessary to account for plant-specific factors. Likewise, for bioremediation uses, final improvement may be necessary in the ultimate host. Many additional herbicide applications of P450 shuffling are found in the

U.S. Patent Application entitled "DNA Shuffling to Produce Herbicide Selective Crops"
Attorney Docket Number 018097-025600US and assigned U.S.S.N. _____.

Table 2 displays homology between selected cytochrome P-450 genes preferred for use in this embodiment of the invention.

5

Table 2. DNA homology between selected cytochrome P450 genes.

CYP105A1 suaC	CYP105D1 soyC	CYP105B1 subC	CYP105A2	Sca2	
--	58%	51	56	78	105A1
	--	51	48	57	105D1
		--	56	52	105B1
			--	53	105A2
				--	Sca2

In addition to these monooxygenase mediated reactions, the use of reactions that are
mediated by polypeptides that do not have monooxygenase activity is also within the scope
of the present invention. In a preferred embodiment, these non-monooxygenase
polypeptides will operate on a substrate that has been acted on by a monooxygenase. In
another preferred embodiment, these polypeptides will operate on a compound prior to its
being acted on by a monooxygenase. Moreover, it is within the scope of the present
invention to improve one or more properties of the non-monooxygenase polypeptides by
shuffling nucleic acids encoding these polypeptides.

C. Accessory Polypeptides

In conjunction with the oxidative pathways utilizing polypeptides having
monooxygenase activity, as discussed above, the present invention provides accessory non-
monooxygenase polypeptides. As used herein, "accessory polypeptides" refers to those
polypeptide that do not carry out the initial monooxidation step in the methods of the
invention. Exemplary accessory polypeptide include, ligases, transferases, dehydrogenases,
and the like. Although both shuffled and non-shuffled polypeptides can be used, preferred
accessory polypeptides are those that have been shuffled.

The non-monooxygenase polypeptides can be used at any step of a pathway of the invention. In a preferred embodiment, they will be used to further transform the oxidation product. Although it will generally be preferred to utilize oxidized substrates that are produced by a monooxygenase of the invention, those of skill will appreciate that these routes can be practiced with analogous substrates that are, for example chemically synthesized, commercially available, *etc.*

Moreover, the present invention provides methods using both the improved accessory peptides and unimproved accessory peptides to further elaborate the monooxygenase-mediated reaction product. The method includes contacting the product of the monooxygenase-mediated reaction with one or more of the accessory polypeptides. In a preferred embodiment, the product is contacted with an organism that expresses the accessory polypeptide(s). When the accessory polypeptides are improved polypeptides, they will generally be produced by the methods described herein.

The improved monooxygenase and the accessory polypeptide(s) can be expressed by the same host cell, or they can be expressed by different host cells. In a preferred embodiment, the accessory polypeptide is an improved polypeptide.

By utilizing accessory polypeptides, the present invention makes possible the synthesis of a great variety of industrially valuable compounds via the methods disclosed herein.

1. *Dehydrogenases*

In a preferred embodiment, an alcohol or diol is converted to an aldehyde or carboxylic acid by the action of a dehydrogenase. The substrate for the dehydrogenase is preferably the product of an improved oxygenase of the invention.

Polynucleotides encoding many known dehydrogenases can be used as substrates for DNA shuffling. Exemplary dehydrogenases useful in practicing the present invention include, but are not limited to:

[ECOALDB, ECAE000436, ECAE000239, D90780, D90781, ECOFUCO, ECOFUCO] dehydrogenase of *Escherichia coli*; [AF029734 and AF029733] dehydrogenase of *Xanthobacter autotrophicus*; [AREXOYGEN] dehydrogenase of *Agrobacterium radiobacter*; [AB003475] dehydrogenase of *Deinococcus radiodurans*; [AF034434, VIBTAGALDA] dehydrogenase of *Vibrio cholerae*; [D32049] dehydrogenase of *Synechococcus* sp.; [AE001154] dehydrogenase of *Borrelia burgdorferi* (BB0528); [ABY17825]

dehydrogenase of *Agaricus bisporus*; [ASNALDAA] dehydrogenase of *Aspergillus niger*; [EMEALDA, EMEALCA] dehydrogenase of *Aspergillus nidulans*; [AF019635, PPU15151] dehydrogenase of *Pseudomonas putida* TOL plasmid, xylW, xyl C; [AF031161] dehydrogenase of *Pseudomonas sp.* VLB120, (stdD); [PFSTYABCD] dehydrogenase of *P. fluorescens*, styD; [PPU24215] dehydrogenase of *P. putida*, F1p-cymene alcohol and aldehyde dehydrogenases.

2. Conversion of hydroxyls and/or acids to esters

In another preferred embodiment, there is provided a method for converting carboxylic acid and hydroxyl groups to adducts such as esters and ethers. Useful polypeptides include, for example, ligases and transferases (see, Fig. 4). For the purposes of the discussion below, these polypeptides are referred to as "adduct-forming" polypeptides.

The adduct-forming polypeptides are useful for enhancing and controlling the production of biotransformation products. These polypeptides, which convert a diol, for example, to a monoacyl or monoglycosyl derivative can enhance control over the regioselectivity of subsequent reactions (e.g., chemical dehydration). For example, the regioselectivity of chemical dehydration in certain cases can be controlled by converting the compounds to their diacyl derivatives by means of chemical reaction, and then selectively removing one of the acyl groups using an polypeptide of the invention. Alternatively, one can control the regioselectivity of the dehydration by using an esterase or a trans-acylase polypeptide to convert the compounds to monoacyl derivatives, preferably in the presence of an excess of another carboxylic acid ester. In addition, the isolation of certain products is simplified by their conversion to more hydrophobic species. For example, the acylation of diols to the corresponding carboxylic ester provides for a more efficient recovery of such diols, in the form of an ester, by organic solvent extraction of the adduct. Preferred organic solvents are those that can be used in an immiscible biphasic organic-aqueous biotransformation with whole cells, whether in a batch or in a continuous mode.

An adduct-forming polypeptide can be expressed by the same host cell that expresses the dioxygenase, dehydrogenase, racemase, etc., or it can be expressed by a different host cell. Moreover, an adduct-forming polypeptide can be a naturally occurring polypeptide, or it can be improved by the method of the invention.

When the adduct-forming polypeptide is an improved polypeptide, in presently preferred embodiments, the polypeptides demonstrates increased efficiency in the formation of the monoacyl- or monoglycosyl- derivatives of a desired compound (*e.g.*, a glycol, carboxylic acid, *etc.*). Other improved adduct-forming polypeptides include

5 transferases and ligases that can selectively modify only one of the hydroxyl groups of a diol, thus providing a means for controlling the regioselectivity of dehydration of such derivatives to either of two possible isomeric α -hydroxycarboxylic acid compounds.

a. Acyltransferases

10 One class of enzymes useful in practicing the present invention are the acyltransferases. These polypeptides can be evolved to enhance certain catalytic properties of the encoded polypeptides such as, specificity for a particular hydroxyl and/or acid, enantiomeric and/or diastereomeric selectivity.

More specifically, these polypeptides catalyze acyl transfer reactions as

15 shown in Fig. 4. Acyltransferases are ubiquitous in nature, and many organisms (*e.g.*, microbes, plants, mammals, *etc.*) can be used as sources of genes encoding these polypeptides. No matter their origin, the acyltransferase genes are preferably selected from those encoding functional polypeptides that catalyze active (CoA) ester transfer reactions in the biocatalytic processes described herein. Preferred acyltransferase genes are selected

20 from those encoding functional polypeptides catalyzing reactions of small non-biopolymeric molecules.

Examples of various acyltransferases useful in the present invention include polypeptides that catalyze the methylation of α -hydroxycarboxylic acids. A list of exemplary polynucleotides that can be recruited for this purpose are listed below by the

25 corresponding GenBank identification:

[AF043464] acetyl-CoA: benzylalcohol acetyltransferase of *Clarkia breweri*, and benzoyl-CoA benzyl alcohol acetyltransferase present in the same organism, (Dudareva *et al*, *Plant Physiol.* 116(2):599-604 (1998));

[DCANTHRAN, DCHCBT1, DCHCBT1A, DCHCBT1B, DCHCBT2, DCHCBT3] hydroxycinnamoyl/benzoyl-CoA:anthranilate N-acyltransferase

30 of *Dianthus caryophyllus*; [E08840] homoserine o-acetyltransferase of *Acremonium chrysogenum*; [E12754] anthocyanin 5-aromatic acyltransferase, of *Gentiana triflora*; [HUMBCAT] branched chain acyltransferase (human,

J03208, J04723); [MG396;D02^oorf152(lacA); MJ1064(lacA) MJ1678, MTH1067]; galactoside 6-O acetyl transferase EC 2.3.1.18, lac A of *E. coli* ; B0342(lacA); or of other organisms; [B3607(cysE), HI0606(cysE), HP1210(cysE), SLR1348(cysE)] serine O-acetyltransferase EC 2.3.1.30; [YGR177C, YOR377W] alcohol O-acetyltransferase, EC 2.3.1.84, of *Saccharomyces cerevisiae*; [e.g., Q00267, D90786, Z92774, I78931 AF030398, AF008204, AF042740] arylamine N-acetyltransferase, EC 2.3.1.118; [YAR035(YAT1), YM8054.01(CAT2)] carnitine O-acetyltransferase, EC 2.3.1.7, or mammalian origin of from yeast; [CHAT] choline O-acetyltransferase, EC 2.3.1.6, of mammalian origin; acetyl CoA:deacetylindoline 4-O-acetyltransferase (EC 2.3.1.107) St-Pierre *et al*, *Plant J.* 14(6): 703-713 (1998); and [ECOPLSC] 1-acyl-sn-glycerol-3-phosphate acyltransferase (plsC) of *Escherichia coli*.

b. Acyl CoA ligases

In another embodiment an accessory polypeptide having acyl CoA ligase activity is provided.

The specificity of acyl-CoA ligases towards a particular exogenous substrate or a group of substrates is preferably optimized by screening or selecting for the acylation of a substrate by shuffled and co-expressed acyl-CoA ligases and acyltransferases. Utilizing these polypeptides in tandem allows the combined effect of both polypeptides to be exploited.

To illustrate the family or single gene shuffling approach to improving acyl-CoA ligases or acyltransferases, one more of the more members of the corresponding superfamilies of these polypeptides are selected, aligned with similar homologous sequences, and shuffled against these homologous sequences.

An exemplary list of useful acyl-CoA ligase genes for inclusion into an organism of the invention is provided below:

[AF029714, ECPAA, AJ000330, PSSTYCATA] phenylacetate-CoA ligase, EC 6.2.1.30; [Y11070, Y11071] phenylpropionate-CoA ligase; [B2260(menE), SLR0492(menE), SAU51132(menE)] O-succinylbenzoate-CoA ligase, EC 6.2.1.26; [RPU75363, RBLBADA, AA532705, AA664442, AA497001, AF042490, ARGFCBABC] (chloro)benzoate-CoA ligase, EC

6.2.1.25; [SBU23787, VPRNACOAL, POTST4C11, RIC4CL2R, OS4CL, AF041051, AF041052, GM4CL14, GM4CL16, LEP4CCOALA, LEP4CCOALB, PC4CL1A, PC4CL1AA, PC4CL2A, PC4CL2AA, TOB4CCAL, TOBTCL2, TOBTCL6, ECO110K, AF008183, AF008184, AF041049, AF041050, ATU18675, NTU5084, NTU50846, PTU12013, PTU39404, PTU39405, ATF13C5, ORU61383, AF064095, AA660600, AA660679, STMPABA] 4-coumarate-CoA ligase EC 6.2.1.12; [RPU02033] 4-hydroxybenzoate-CoA ligase; [PSPPLAS] 2-aminobenzoate-CoA ligase.

In some embodiments of the invention, a carboxylic acid is fed exogenously to the organism that expresses the ligase or transferase. Preferably, the carboxylic acid is selected from those compounds that cannot be altered by the polypeptide used to produce the substrate acted upon by the adduct forming polypeptide. Such carboxylic acids include, for example, both substituted and non-substituted benzoic acid, phenylacetic acid, naphthoic, phenylpropionic acid, phenoxyacetic acid, cycloalkanoic acid, carboxylic acids derived from terpenes, pivalic acid, substituted acrylic acids, and the like.

To facilitate the utilization of exogenously supplied carboxylic acids, and for enhancing the variety of compounds suitable for use in this process, the invention also provides microorganisms in which one or more mutations are introduced. Preferred mutations are those that effectively block metabolic modifications of such acids beyond their conversion to a suitable active ester (*e.g.*, as a derivative of coenzyme A). Such mutations in the host organism can be introduced by classical mutagenesis methods, by site-directed mutagenesis, by whole genome shuffling, and other methods known to those of skill in the art. One can also introduce mutations that minimize host endogenous esterase activity.

In a presently preferred embodiment, the acyl transferase-encoding nucleic acids used as substrates for creating recombinant libraries encode polypeptides that transfer an acetyl group from an endogenous pool of acetyl-CoA in the cells of the host. The endogenous pools of acetyl-CoA can also be enhanced by DNA shuffling of an acetyl-CoA ligase and by supplying an exogenous acetate in the medium.

While using acetyl-CoA transferases or other acyltransferase or glycosyltransferase does not necessarily require expression of a corresponding acetyl-CoA or other ligase, in a presently preferred embodiment, the organisms produce a sufficient amount of an acyl-CoA ligase so as to activate the carboxylic acids to CoA thioesters, which in turn serve as substrates for acyl-CoA transferases that utilize the oxidation products as substrates.

The specificity of an acyl-CoA ligase towards a desired exogenous carboxylic acid can be optimized using the recombination and screening/selection methods of the invention. Preferably, the screening or selecting is performed using co-expressed acyl-CoA ligases and acyltransferases, thus permitting one to screen on the basis of the combined effect of both polypeptides in the pathway for provision of monoacylated derivatives of the oxidation products.

Nucleic acids that encode acyl-CoA ligases and other acyltransferases useful as substrates for the recombination and selection/screening methods of the invention include, for example, one or more members of the superfamilies of these polypeptides. In a presently preferred embodiment, the nucleic acids are selected, aligned with similar homologous sequences, and shuffled against these homologous sequences.

c. Glycosyltransferases

Similarly, one or more glycosyltransferases can be expressed by the host cells of the invention. Alternatively, one or more glycosyltransferases can be selected from the glycosyltransferase superfamily, aligned with similar homologous sequences, and shuffled against these homologous sequences. Glycosyl transfer reactions are ubiquitous in nature, and one of skill in the art can isolate such genes from a variety of organisms, using one or more of several art-recognized methods. The following are illustrative examples of glycosyltransferase-encoding nucleic acids that can be used as substrates for creation of the recombinant libraries. The libraries are then screened to identify those polypeptides that exhibit an improvement in the glycosylation of compounds such as alcohols, diols and α -hydroxycarboxylic acids:

[EC 2.4.1.123] inositol 1- α -galactosyltransferase; [NTU32643, NTU32644]
phenol β -glucosyltransferase, EC 2.4.1.35; flavone 7-O-beta-glucosyltransferase, EC 2.4.1.81; [AB002818, ZMMCCBZ1, AF000372, AF028237, AF078079, D85186, ZMMC2BZ1, VVUFGT]; flavonol 3-O-glucosyltransferase, EC 2.4.1.91; o-dihydroxycoumarin 7-O-glucosyltransferase, EC 2.4.1.104; vitexin beta-glucosyltransferase, EC 2.4.1.105; coniferyl-alcohol glucosyltransferase, EC 2.4.1.111; monoterpene beta-glucosyltransferase, EC 2.4.1.127; arylamine glucosyltransferase, EC 2.4.1.71; sn-glycerol-3-phosphate 1-galactosyltransferase, EC 2.4.1.96; [RNUDPGTR, AA912188, AA932333] glucuronosyltransferase, EC 2.4.1.17;

the human UGT and isoenzymes (~35 genes); salicyl-alcohol glucosyltransferase, EC 2.4.1.172; 4-hydroxybenzoate 4-O-beta-D-glucosyltransferase, EC 2.4.1.194; zeatin O-beta-D-glucosyltransferase, EC 2.4.1.203; [VFAUDPGFTA] D-fructose-2-glucosyltransferase; and [MBU41999] ecdysteroid UDP-glucosyltransferase (egt).

In presently preferred embodiments, the glycosyltransferases are selected from those which transfer hexose residues from UDP-hexose derivatives. Preferred hexoses include, for example, D-glucose, D-galactose and D-N-acetylglucosamine.

d. Methyltransferases

In a still further preferred embodiment, the host cells of the present invention express a polypeptide capable of converting a carboxylic acid to a carboxylic acid methyl ester. Presently preferred polypeptides include methyltransferases.

For the purpose of this invention, genes encoding S-adenosylmethionine-dependent methyltransferases are preferred. In a preferred embodiment, these polypeptides are evolved to enhance selected properties of the encoded polypeptides such as, specificity for a particular substrate and enantiomeric and/or diastereomeric selectivity and/or solvent resistance.

More specifically, these polypeptides can be evolved to catalyze the O-methylation of carboxyl groups of a carboxylic acid substrate thus forming the corresponding methyl esters. Methyltransferases are ubiquitous in nature, and many organisms (*e.g.*, microbes, plants, mammals, *etc.*) can be used as sources of genes encoding these polypeptides. No matter their origin, the methyltransferase genes are preferably selected from those which encode functional polypeptides that catalyze the methylation of small non-biopolymeric molecules. Preferably, the methyltransferases are those which act on the carboxyl groups of organic acids.

Examples of various methyltransferases that can be expressed by host cells of the invention and which are useful for DNA shuffling-based directed evolution of polypeptides catalyzing the methylation of carboxylic acids are listed below by the corresponding GenBank identification:

[SCCCAGC3] methyltransferase of *Streptomyces clavuligerus*
methyltransferase CmcJ; [SEERYGENE] methyltransferase of *S. erythraea*
methyltransferases; [SEU77454] methyltransferase of *Saccharopolyspora*

5 *erythraea*; erythromycin O-methyltransferase (eryG); [SGY08763]
 methyltransferase of *S.griseus*; [SKZ86111] methyltransferase of *S.lividans*;
 [STMDNRDKP] methyltransferase of *Streptomyces peucetius*; carminomycin
 o-methyltransferase (dnrK); [MDAJ39670] methyltransferase of
 10 *Streptomyces ambofaciens*; [SEY14332] methyltransferase of
Saccharopolyspora erythraea; [SPU10405] methyltransferase of
Streptomyces purpurascens ATCC 25489; [STMDAUA] methyltransferase of
Streptomyces sp.; aklanonic acid methyltransferase (dauC), and
 carminomycin 4-O-methyltransferase (dauK); [SC2A11 and SC3F7]
 15 methyltransferase of *Streptomyces coelicolor*; [SHGCPIR] methyltransferase
 of *S.hygroscopicus*; [STMCARMETH] methyltransferase of *Streptomyces*
peucetius carminomycin 4-O-methyltransferase; [STMODPOMT]
 methyltransferase of *Streptomyces alboniger* O-demethylpuromycin-O-
 methyltransferase (dmpM); [STMTCREP]; methyltransferase of
 20 *Streptomyces glaucescens*; [SLLMRBG] methyltransferase of *S. lincolnensis*
 lmrB methyltransferase; [SSU65940] 31-O-demethyl-FK506
 methyltransferase (fkbM) of *Streptomyces* sp.; [STMDAUABCE] aklanonic
 acid methyltransferase (dauC) of *Streptomyces* sp.; [STMMDMBC] O-
 methyltransferase (mdmC) of *Streptomyces mycarofaciens*; [STMTYLF]
 25 macrocyn-O-methyltransferase (tylF) of *S.fradiae*; [E08176] Gene of
 mycinamicin III-O-methyltransferase; [AF040571] methyltransferase of
Amycolatopsis mediterranei; [ECU56082] S-adenosylmethionine:2-
 demethylmenaquinone methyltransferase (menG) of *Escherichia coli*;
 [RHANODABC] methyltransferase (nodS) of *Azorhizobium caulinodans*;
 30 [YSCSTE14] isoprenylcysteine carboxyl methyltransferase (STE14) of
Saccharomyces cerevisiae; [YSCMTSW] farnesyl cysteinecarboxyl-
 methyltransferase (STE14) of *Saccharomyces cerevisiae*; [YSCDHHBMET]
 3,4-dihydroxy-5-hexaprenylbenzoate methyltransferase (COQ3) of
S.cerevisiae; [AF004112 and AF004113] phospholipid methyltransferases
 (cho1+), (cho2+) of *Schizosaccharomyces pombe*; [ASNOMT, ASNOMT1A,
 ASNOMT1B, ASNOMT1C and AF036808-AF036830] O-methyltransferases
 of *Aspergillus*; [MSU20736] S-adenosyl-L-methionine; trans-caffeoyl-CoA3-
 O-methyltransferase of *Medicago sativa*; [ALFIOM] isoliquiritigenin 2'-O-

methyltransferase of *Medicago sativa*; [MSU20736] S-adenosyl-L-methionine; trans-caffeoyl-CoA 3-O-methyltransferase (CCOMT) of *Medicago sativa*; [MSAF000975] 7-O-methyltransferase (7-IOMT(6)) of *Medicago sativa*; [MSAF000976] 7-C-methyltransferase (7-IOMT(9)) of *Medicago sativa*; [MSU97125] of isoflavone-O-methyltransferase *Medicago sativa*; [NTCCOAOMT] caffeoyl-CoA O-methyltransferase of *Nicotiana tabacum*; [NTZ82982] caffeoyl-CoA O-methyltransferase 5 of *N. tabacum*; [NTDIMET] o-diphenol-O-methyltransferase of *N. tabacum*; [PCCCOAMTR, PUMCCOAMT] trans-caffeoyl-CoA 3-O-methyltransferase of *Petroselinum crispum*; [PTOMT1] s caffeic acid/5-hydroxyferulic acid O-methyltransferase (PTOMT1) of *Populus tremuloide*; [PBJAJ4894-PBJAJ4896] caffeoyl-CoA 3-O-methyltransferases of *Populus balsamifera* subsp. *trichocarpa*; [ZEU19911] S-adenosyl-L-methionine: caffeic acid 3-O-methyltransferase of *Zinnia elegans*; [SLASADEN] S-adenosyl-L-methionine:trans-caffeoyl-CoA 3-O-methyltransferase of *Stellaria longipes*; [VVCCOAOMT] caffeoyl-CoA O-methyltransferase of *V. vinifera*; [D88742] O-methyltransferase of *Glycyrrhiza echinata*; [AF046122] caffeoyl-CoA 3-O-methyltransferase (CCOMT) of *Eucalyptus globulus*; [ATCOQ3] dihydroxypolyprenylbenzoate: methyltransferase of *Arabidopsis thaliana* [CSJSALMS90] S-adenosyl-L-methionine:scoulerine 9-O-methyltransferase of *Coptis japonica*; [HVU54767] caffeic acid O-methyltransferase (HvCOMT) of *Hordeum vulgare*; [MCU63634] inositol methyltransferase (Imt1) of *Mesembryanthemum crystallinum*; [PSU69554] 6a-hydroxymaackiain methyltransferase (hmm6) of *Pisum sativum*; [CAU83789] O-diphenol-O-methyltransferase of *Capsicum annum*; [U16794] 3' flavonoid O-methyltransferase (form. 1) of *Chrysosplenium americanum*; [CBU86760] SAM:(Iso)eugenol O-methyltransferase (IEMT1) of *Clarkia breweri*; salicylic acid carboxyl SAM-O-methyltransferase (Dudareva *et al*, *Plant Physiol.* 116(2):599-604 (1998)); [HSHIOMT9] hydroxyindole-O-methyltransferase (HIOMT) of *Homo sapiens*; [HSCOMT2] gene catechol O-methyltransferase of *Homo sapiens*; [HUMPNTMTA] phenylethanolamine N-methyltransferase gene of *Homo sapiens*; [HUMCOMTA] catechol-O-methyltransferase of *Homo sapiens*; [HUMCOMTC] catechol-O-methyltransferase of *Homo*

sapiens; [HUMPNT] phenylethanolamine N-methyltransferase of *Homo sapiens*; [AF064084] prenylcysteine carboxyl methyltransferase (PCCMT) of *Homo sapiens*; [HUMCMT] carboxyl methyltransferase of *Homo sapiens*; [HUMHNMA] histamine N-methyltransferase of *Homo sapiens*; [RATCATAA, RATCATAB] catechol-O-methyltransferase of *R. norvegicus*; [RATDHNPBMT] dihydroxypolyprenylbenzoate methyltransferase of *Rattus norvegicus*; [BOVPNTB] of Bovine phenylethanolamine N-methyltransferase; [MPEMT7] phosphatidylethanolamine-N-methyltransferase of *Mus musculus* 2; [MMU86108] nicotinamide N-methyltransferase (NNMT) of *Mus musculus*; [MUSCMT] carboxyl methyltransferase protein of Mouse; [GDHOMT] hydroxyindole-O-methyltransferase of *G. domesticus*; [DRU37434] L-isoaspartate (D-aspartate) O-methyltransferase (PCMT) of *Danio rerio*; [DMU37432] protein D-aspartyl, L-isoaspartylmethyltransferase of *Drosophila melanogaster*; and [HAU25845 and HAU25846] farnesoic acid o-methyl-transferases of *Homarus americanus*.

3. Epoxide hydrolases

In a still further preferred embodiment, the present invention provides a nucleic acid encoding a polypeptide capable of converting a particular epoxide to the corresponding diol.

Presently preferred polypeptides include epoxide hydrolases. Many epoxide hydrolases are known, and these enzymes have various substrate specificity and enantioselectivity. Examples of prokaryotic genes encoding epoxide hydrolases suitable for effecting epoxide hydrolysis relevant to this invention include, but are not limited to, [CAJ4332] *Corynebacterium* sp.; and [ARECHA] *Agrobacterium radiobacter* (echA).

In a presently preferred embodiment, the polypeptide has one or more improved properties brought about by shuffling methods described herein. Thus, the nucleic acids encoding this gene, and any homologs of thereof, are subjected to DNA shuffling to evolve polypeptides having improved or optimal performance and specificity towards particular substrates such as α -hydroxycarboxylic acids. In a preferred embodiment, the polypeptide has a performance and/or specificity that is enhanced over the wild type.

Preferred polypeptides act on α -hydroxycarboxylic acid substrates, such as those displayed in Fig. 3.

4. *Enantiomeric interconversion.*

5 In a still further preferred embodiment, the present invention provides a nucleic acid encoding a polypeptide capable of converting a particular enantiomer of a chiral compound such as an alcohol, diol or α -hydroxycarboxylic acid or a precursor or analogue thereof to its antipode.

Presently preferred polypeptides include racemases, such as the mandelate
10 racemase of *Pseudomonas putida* (PSEMDLABC). These polypeptides can be expressed by hosts of the invention in their natural form or, alternatively, they can be evolved to enhance certain catalytic properties of the encoded polypeptides such as, specificity for a particular substrate and enantiomeric and/or diastereomeric selectivity.

The nucleic acids encoding the mandelate racemase of *Pseudomonas putida*,
15 which catalyzes the interconversion of mandelate R and S enantiomers, is a typical preferred example of genes selected for use in this invention. The nucleic acids encoding this gene, and any homologs of thereof, are subjected to DNA shuffling to evolve polypeptides having improved or optimal performance and specificity towards particular substrates such as α -hydroxycarboxylic acids. In a preferred embodiment, the polypeptide has a performance
20 and/or specificity that is enhanced over the wild type. Preferred polypeptides act on α -hydroxycarboxylic acid substrates, such as those displayed in Fig. 3.

5. *α -Ketocarboxylic acid decarboxylase*

Several thiamine phosphate-dependent polypeptides of this class are known to
25 occur in bacteria, fungi and yeast (*see*, Iding et al., Biochim. Biophys. Acta 1358:307-22 (1998)). For the purpose of illustration, a gene encoding a well-known decarboxylase, preferably a benzoylformate decarboxylase (*mdlC*) of *Pseudomonas putida* [PSEMDLABC], is shuffled to increase the specific activity towards α -ketocarboxylic acids, such as o-hydroxybenzalpyruvate. Alternatively, genes encoding pyruvate decarboxylases (EC
30 4.1.1.1), indole-3-pyruvate decarboxylases (EC 4.1.1.74) or phenylpyruvate decarboxylases (EC 4.1.1.43) from a variety of sources can be used.

6. Solvent resistance polypeptides

The invention also provides organisms expressing one or more of the improved polypeptides of the invention and that are also resistant to solvents, organic substrates and reaction products (*e.g.*, epoxides, glycols, α -hydroxyaldehydes, α -hydroxycarboxylic acids and α -hydroxycarboxylic acid derivatives (*e.g.*, esters)) according to the methods of the invention.

The solvent resistance of organisms and polypeptide used in the biocatalytic conversion of organic compounds is important for enhancing the productivity of such processes. Increased solvent resistance of the organisms can enhance longevity, viability and catalytic activity of the microbial cells, and can simplify the administration of the feedstock compounds to the reactor and the recovery or separation of desired products by means of, for example, continuous or semi-continuous liquid-liquid extraction.

In another aspect, the invention provides microbial cells that are useful in the synthetic methods described herein, which express proteins conferring resistance to solvents (in particular, organic solvents) upon the microbial cells. This allows the use of whole microbial cells in a organic-aqueous mixture (*e.g.*, a biphasic mixture). In presently preferred embodiments, the invention provides microbial strains including at least two of the polypeptide systems described herein. For example, a microorganism of the invention can contain both a dioxygenase gene and a transferase gene. In other embodiments, the microorganism can contain both an arene dioxygenase gene and a solvent resistance gene. The microbial cells thus provide a significant improvement in productivity of the synthesis processes, selectivity of product formation, operational simplicity, ease of product recovery and minimizing any by-product streams.

Several microorganisms are known to possess high resistance to hydrophobic compounds such as benzene and lower alkylbenzenes. Recently, genes encoding a solvent efflux pump (*srpABC*) have been identified in *Pseudomonas putida* strains (Kieboom *et al. J. Biol. Chem.* 273:85-91 (1998)). Similarly, various genes that encode polypeptides that confer organic solvent resistance can be found in bacterial strains such as *Pseudomonas putida* GM73 (Kim *et al. J. Bacteriol.* 180: 3692-3696 (1998)), *Pseudomonas putida* DOT-T1E (Ramos *et al. J. Bacteriol.* 180: 3323-3329 (1998)), *Pseudomonas idaho* (Pinkart and White *J. Bacteriol.* 179: 4219-4226 (1997)). These and other genes, such as those that encode many proton-dependent multidrug efflux systems, *e.g.*, MexA-MexB-OprM, MexC-

MexD-OprJ, and MexE-MexF-OprN of *Pseudomonas aeruginosa* (Li *et al.* *J. Bacteriol.* **180**: 2987-2991 (1998)), or the *tolC*, *acrAB*, *marA*, *soxS*, and *robA* loci of *Escherichia coli* (Aono *et al.*, *J. Bacteriol.* **180**:938-944 (1998); White *et al.*, *J. Bacteriol.* **179**:6122-6126 (1997)), and in many other microorganisms, can be used to confer solvent resistance upon a host microbial strain used in the oxidative biocatalytic conversion of olefins by means of action of dioxygenases or dioxygenases.

In presently preferred embodiments, the ability of a polypeptide to confer solvent resistance is enhanced by subjecting nucleic acids encoding solvent resistance polypeptides, or the genomes of the microorganisms themselves, to the recombination and selection/screening methods described herein. The nucleic acids listed above, as well as similar genes, provide a source of substrates for incorporation into organisms of the invention and/or use in DNA shuffling and other methods of constructing libraries of recombinant polynucleotides. The libraries can then be screened to identify those nucleic acids that encode polypeptides conferring improved solvent tolerance on a host. For example, one can select for improved tolerance to compounds such as olefins, AHAs, aldehydes, esters and hydrophobic solvents, including alkanes, cycloalkanes, alcohols and halocarbon derivatives, for example, which are used for performing biotransformation (*e.g.*, two-phase oxidation) of olefins to glycols, AHAs and to their corresponding acyl- and glycosyl- derivatives, *etc.* Similarly, DNA shuffling of nucleic acids that encode these polypeptides can be used to confer and to improve resistance of the microbial cell to high concentrations of biotransformation substrates, intermediates and endproducts, thus improving biocatalyst performance and productivity.

In addition to each of the methods set forth above, the present invention provides polypeptides produced according to these disclosed methods. Moreover, the invention provides organisms that express the polypeptides produced by the method of the invention. The organisms of the invention can express one or more of the improved polypeptides. Also provided by the present invention are methods of synthesizing a desired compound. This method includes contacting an appropriate substrate with a polypeptide of the invention. In a preferred embodiment, the substrate is contacted with an organism of the invention that expresses a polypeptide of the invention.

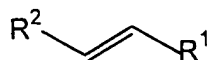
D. Methods of Using Improved Polypeptides to Prepare Organic Compounds

In addition to the methods discussed above, the present invention provides a range of methods for preparing useful organic compounds by the oxidation and further elaboration of appropriate precursors. Among the methods provided by the present invention are, for example, the oxidation of alkylarene compounds to the corresponding unsaturated diols and the subsequent dehydration of these diols hydroxy alkylarenes. Additionally, there is provided an analogous method for preparing hydroxylated aromatic carboxylic acids. Moreover, the invention provides methods for preparing cyclic exocyclic and/or acyclic diols from molecules having alkene bonds. The exocyclic and acyclic diols can be readily converted to α -hydroxycarboxylic acids.

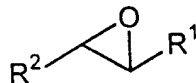
The reaction types and sequences set forth below are illustrative of the scope of the invention. The monooxygenases of the invention are capable of oxidizing any organic substrate comprising an oxidizable moiety. Additional reaction sequences utilizing the polypeptides of the invention will be apparent to those of skill in the art.

1. *Preparation of epoxides*

In a preferred embodiment, there is provided a method for converting an olefin into an epoxide. The polypeptide of the invention is designed to be functional with substantially any olefinic substrate, however, in a preferred embodiment, the polypeptide acts on at least one alkene group of a substrate that includes:



to produce an epoxide product having the structure:



wherein, R^1 and R^2 are independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl, substituted heterocyclyl, $\text{---NR}^3\text{R}^4(\text{R}^5)_m$, ---OR^3 , ---CN , $\text{C}(\text{R}^6)\text{NR}^3\text{R}^4$ and $\text{C}(\text{R}^6)\text{OR}^3$ groups. R^3 , R^4 and R^5 are members independently selected from the group consisting of H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl and substituted

heterocyclyl groups. R^5 is selected from $=O$ and $=S$. m is 0 or 1, such that when m is 1, an ammonium salt is provided.

In a still further preferred embodiment, the olefinic substrate is selected from 2-vinylpyridine, 4-vinylpyridine, 3-butenitrile, vinylacetamide, N,N-dialkyl vinylacetamide, diallylamine, triallylamine, diallyldimethylammonium salts, styrene and phenyl-substituted styrene.

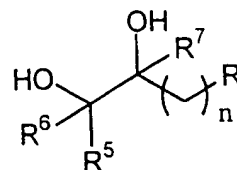
2. Preparation of vicinal diols

The formation of vicinal diols by oxidizing a π -bond using a monooxygenase of the invention and hydrolyzing the resulting epoxide provides ready access to a wide array of compounds that are useful as both final products and as intermediates in multi-step reaction pathways. The monooxygenases of the invention are capable of converting to expoxides and, thus, to vicinal diols an array of structurally distinct compounds comprising one or more π -bonds.

Although the method can be practiced with essentially any π -bond, in essentially any compound, in a preferred embodiment, the method includes preparing a vicinal diol group by contacting a substrate comprising a carbon-carbon double bond with an improved monooxygenase polypeptide, or an organism expressing an improved monooxygenase polypeptide to form an epoxide. The epoxides are cleaved by chemical or enzymatic action.

In another preferred embodiment, the substrate comprising the carbon-carbon π -bond is selected from styrene, substituted styrene, divinylbenzene, substituted divinylbenzene, isoprene, butadiene, diallyl ether, allyl phenyl ether, substituted allyl phenyl ether, allyl alkyl ether, allyl aralkyl ether, vinylcyclohexene, vinylnorbornene, and acrolein.

In yet another preferred embodiment, the vicinal diol produced by the action of the improved monooxygenase polypeptide has the structure:



wherein R^1 and R^5 are independently selected from alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl, substituted heterocyclyl, $-NR^2R^3$, $-OR^2$, $-CN$, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups, or R^1 and R^5 are joined to form a ring

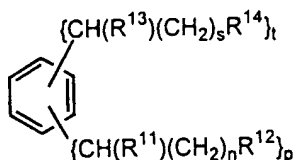
system selected from saturated hydrocarbyl rings, unsaturated hydrocarbyl rings, saturated heterocyclyl rings and unsaturated heterocyclyl rings; R^2 and R^3 are members independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl and substituted heterocyclyl groups; R^4 is selected from =O and =S; R^6 and R^7 are independently selected from H and alkyl; and n is a number from 0 to 10, inclusive.

In certain preferred vicinal diols R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl $—NR^2R^3$, $—OR^2$, $—CN$, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups, R^2 and R^3 are members independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl and substituted heterocyclyl groups; and R^4 is selected from =O and =S.

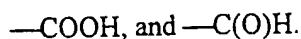
In another preferred embodiment, the diol includes a six-member ring having at least one endocyclic double bond and at least one substituent selected from methyl, carboxyl and combinations thereof.

3. Dehydrogenation of ROH groups

In an other preferred embodiment, the invention provides a class of improved P-450 polypeptides that dehydrogenate hydroxyl-containing substrates. Although substantially any hydroxyl-containing substrate can be dehydrogenated using the polypeptides of the invention, in a preferred embodiment, the substrate is:



wherein R^{11} , R^{12} , R^{13} and R^{14} are independently selected from H and OH and at least one of R^{11} , R^{12} , R^{13} and R^{14} is OH; n and s are independently selected from the numbers 0 to 16; and p and t are independently selected from 0 to 6, wherein at least one of p and t must be at least one. The enzyme of the invention, preferably, converts at least one hydroxyalkyl group to a member selected from:



In another preferred embodiment, the substrate is selected from among toluene and xylene and the polypeptide converts said at least one methyl group to a carboxylic acid or a carbonyl.

5 4. *Preparation of α -hydroxycarboxylic acids*

In another preferred embodiment, there is provided a method for converting an olefin to an α -hydroxyaldehyde or an α -hydroxycarboxylic acid. In a preferred embodiment, the olefin is converted to an α -hydroxycarboxylic acid. The method includes: (a) contacting the olefin with an improved monooxygenase polypeptide of the invention to
10 form an epoxide; (b) hydrolyzing the epoxide to form a vicinal diol; and (c) contacting the vicinal diol with a dehydrogenase polypeptide to form the α -hydroxycarboxylic acid.

As in other methods involving the hydrolysis of the epoxide, the epoxide can be hydrolyzed using chemical or enzymatic means. The hydrolysis is preferably mediated by an improved epoxide hydrolase prepared using the methods of the invention. The
15 dehydrogenase polypeptides useful in this embodiment can be naturally occurring polypeptides or, alternatively, they can be polypeptides improved using the methods of the invention. When more than one polypeptide is used to effect a particular transformation they can be expressed in the same host organism or in different host organisms.

α -Hydroxycarboxylic acids (AHAs) are an important group of industrial
20 chemicals. One of the simplest representatives of this class of compounds is lactic acid. Lactic acid is used for many purposes, including the synthesis of polyester polymers (e.g., polylactic acid). In addition to the lactic acid homopolymer, lactic acid can be copolymerized with other α -hydroxycarboxylic acids, such as mandelic acid, to form copolymers with lactic acid. Enantiomerically pure hydroxycarboxylic acids are also used as
25 resolving reagents for separating mixtures of chiral molecules. α -Hydroxycarboxylic acids are generated chemically by a variety of general methods that are less than ideal. For example, a commonly used method, hydrolysis of a cyanohydrin is problematic. The cyanohydrins are produced by the addition of HCN to an aldehyde. Aldehydes are relatively expensive starting materials and the hydrolysis of the cyanohydrins to the corresponding α -
30 hydroxycarboxylic acids does not proceed in an enantioselective manner. This necessitates the disposal or recycling of a substantial portion of the costly aldehydes.

Chiral lactic acid has been manufactured by means of a microbial fermentative process using a carbohydrate feedstock. At present, this fermentative

methodology does not provide a means for making AHAs other than lactic acid. A great number of useful AHAs have a structure wherein the lactic acid methyl group is replaced with another substituent such as, for example, aromatic, alicyclic or alkenic moieties useful for subsequent chemical modifications of either the AHAs themselves, or of polymers or copolymers incorporating these AHAs.

A promising route to the highly selective manufacture of chiral AHAs is based on the oxidation of olefins by means of a monooxygenase polypeptide of the invention. These polypeptides can be isolated and used *in vitro* or, alternatively, they can be used *in vivo* by using whole microbial cells displaying the appropriate polypeptide activity. Moreover, dioxygenase polypeptides also have useful activity. The preparation of α -hydroxy carboxylic acids utilizing dioxygenases is disclosed in U.S.S.N. _____, bearing Attorney Docket No. 018097-031100, entitled "Shuffling of Dioxygenase Genes for Production of Industrial Chemicals", filed on an even date herewith and incorporated by reference in its entirety.

The present invention also provides improved polypeptides that exhibit an enhanced ability to convert a range of substrates to α -hydroxycarboxylic acids, α -hydroxycarboxylic acid precursors and analogues by processes employing oxidative biocatalysis. Methods are provided for generating polynucleotides that encode enzymes that catalyze these reactions and that have improved properties. Presently preferred substrates include olefins.

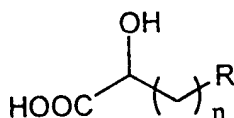
Biocatalytic methods that employ the recombinant polypeptides provided by the present invention have several significant advantages over previously available methods for the synthesis of α -hydroxy acids, their precursors and analogues. For example, the invention provides polypeptides that can increase the amount of product produced in a reaction, as well as increase the enantiomeric excess and/or regiospecific formation of the product. Among the enhanced properties that are obtained using the methods include enhanced forward rate kinetics, altered substrate specificity and affinity, enhanced regioselectivity and enantioselectivity, and decreased susceptibility to inhibitors and inactivation by substrates, intermediates and products.

As is generally true for the other aspects and embodiments of the present invention, the recombinant polypeptides of the invention are preferably expressed by an organism, such as microbial cells, that carry out the biocatalysis. Accordingly, the invention also provides organisms that are adapted for efficient biocatalytic manufacturing of α -

hydroxycarboxylic acids, their analogues and their precursors. The microorganisms preferably express one or more recombinant polypeptides that are optimized for the biocatalysis pathway of interest. The biocatalytic polypeptides that are expressed by the microbial cells can be wild type or they can be recombinant polypeptides that exhibit improved properties encoded by the recombinant nucleic acids obtained using the methods of the invention. In a preferred embodiment, the organism expresses at least two enzymes selected from an improved monooxygenase, an epoxide hydrolase and a dehydrogenase. Either or both of the epoxide hydrolase and the dehydrogenase can be an improved polypeptide.

In yet another embodiment, a nucleic acid encoding a polypeptide that converts a vicinal glycol to an α -hydroxyaldehyde and/or an α -hydroxycarboxylic acid is provided. For the purpose of this invention, the genes encoding dehydrogenase polypeptides for conversion of the glycols to α -hydroxyaldehydes and/or to α -hydroxycarboxylic acids, can be selected from many known dehydrogenases.

In another preferred embodiment, the method of invention is used to convert olefinic and vicinal diol precursors to α -hydroxycarboxylic acids having the structure:



wherein,

R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl, substituted heterocyclyl, $-\text{NR}^2\text{R}^3$, $-\text{OR}^2$, $-\text{CN}$, $\text{C}(\text{R}^4)\text{NR}^2\text{R}^3$ and $\text{C}(\text{R}^4)\text{OR}^2$ groups; R^2 and R^3 are members independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl and substituted heterocyclyl groups; R^4 is selected from $=\text{O}$ and $=\text{S}$, and n is a number between 0 and 10, inclusive.

In a still further preferred embodiment, R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl $-\text{NR}^2\text{R}^3$, $-\text{OR}^2$, $-\text{CN}$, $\text{C}(\text{R}^4)\text{NR}^2\text{R}^3$ and $\text{C}(\text{R}^4)\text{OR}^2$ groups; R^2 and R^3 are members independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl and substituted heterocyclyl groups; and R^4 is selected from $=\text{O}$ and $=\text{S}$.

In yet another preferred embodiment, the invention provides a method for altering or controlling the regiospecificity of the dehydrogenation reaction. This method

“blocks” one of the vicinal diol hydroxyl groups by forming an ester, for example. The method includes contacting the vicinal diol with a microorganism comprising an improved polypeptide having an activity selected from ligase, transferase and combinations thereof, thereby forming a α -hydroxycarboxylic acid adduct. As with the other polypeptides discussed above, this polypeptide can be expressed by the same host cell that expresses other polypeptides of the reaction cascade. Moreover, this polypeptide can be a naturally occurring polypeptide, or it can be improved using the method of the invention.

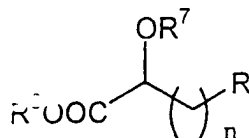
a. α -Hydroxycarboxylic acid adducts

AHAs are bifunctional molecules with two chemically and enzymatically distinguishable functional groups, carboxyl and hydroxyl. In the biocatalytic modifications of AHAs described in this invention, either of these groups can be derivatized by bond formation. While these reactions do not change the oxidation state of the AHA molecule, recruitment of the enzymes effecting modification of AHAs provides the opportunity to generate biotransformation endproducts with substantially different physical and chemical properties than that of a free AHA. Generally desirable properties include an increase of hydrophobicity, a decrease of aqueous solubility and, for an ester formed through a carboxylic group of an AHA, a decrease in acidity of the process end-products.

In a preferred embodiment, the adduct-forming polypeptide produces an α -hydroxycarboxylic acid adduct selected from esters and ethers. The method includes contacting an α -hydroxycarboxylic acid with a polypeptide having an activity selected from ligase, transferase and combinations thereof, thereby forming a α -hydroxycarboxylic acid adduct. The adduct forming polypeptides useful in this embodiment can be naturally occurring polypeptides or, alternatively, they can be polypeptides improved using the methods of the invention, as discussed generally, above.

Exemplary adduct forming reactions are provided in Fig. 4. This Figure shows the use of a methyltransferase to convert carboxylic acid (X) to the corresponding methyl ester (XI), acyltransferase I to convert the X to ester XIII, and acyl-CoA ligase to convert X to intermediate XIV. This intermediate can then be transformed into a simple alkyl ester (XIX) or to structures having greater complexity of structure in the alcohol-derived component (*e.g.*, XV). Species such as XV can be further elaborated using other polypeptides including, for example, acyltransferase III to produce compound XVII, thioesterase II to produce compound XVIII and thioesterase I to produce compound XVI.

In a further preferred embodiment, the α -hydroxycarboxylic acid adduct has the structure:



wherein, R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl, substituted heterocyclyl, $-\text{NR}^2\text{R}^3(\text{R}^4)_m$, $-\text{OR}^2$, $-\text{CN}$, $\text{C}(\text{R}^5)\text{NR}^2\text{R}^3$ and $\text{C}(\text{R}^5)\text{OR}^2$ groups, R^2 , R^3 and R^4 are members independently selected from the group consisting of H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl and substituted heterocyclyl groups; R^5 is selected from $=\text{O}$ and $=\text{S}$; R^6 is selected from H, alkyl and substituted alkyl groups; R^7 is $\text{C}(\text{O})\text{R}^8$, wherein R^8 is selected from H alkyl and substituted alkyl groups and R^7 and R^8 are not both H; m is 0 or 1, such that when m is 1, an ammonium salt is provided; and n is a number between 0 and 10, inclusive.

In yet another preferred embodiment, R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl — NR^2R^3 , — OR^2 , — CN , $C(R^5)NR^2R^3$ and $C(R^5)OR^2$ groups; R^2 and R^3 are members independently selected from the group consisting of H, C_1 - C_6 alkyl and allyl; and R^5 is =O.

In yet another preferred embodiment of this invention, the described reactions and pathways are utilized for biocatalytic whole-cell conversion of styrene to mandelic acid and its ester derivatives. The pathway for styrene conversion, all of its intermediates and reactions are shown in Fig. 2.

The esterified adducts provide an increase in the overall efficiency of the biotransformation process as they simplify end-product recovery. The esters are easily isolated by organic solvent extraction and partitioning. Moreover, the adducts obviate the need for pH adjustment in the aqueous fermentation media to prevent the accumulation of the high levels of acidic biotransformation products.

There are several biochemically distinct means by which AHAs can be biocatalytically esterified in a substantially aqueous environment. In one preferred embodiment of this invention, expression of genes encoding an S-adenosylmethionine (SAM)-dependent O-methyltransferase is used to effect conversion of AHAs to their methyl esters (e.g., **Fig. 4**, conversion of compound X to compound XI). SAM-dependent

methyltransferases of differing substrate specificity are common in nature, and suitable enzymes and corresponding genes can be found and used directly for the purpose of this invention. Alternatively, these species can be further evolved and optimized for specific activity with the AHAs using one or more nucleic acid shuffling methods described herein.

- 5 The invention also provides means for HTP screening for the presence, and quantitative determination, of the AHA-specific O-methyltransferase catalytic activities in microorganisms, cells, tissues or extracts of tissues of higher eukaryotic organisms. These methods can be used either to identify sources of corresponding genes or to evolve the desired specificity of known methyltransferases towards the AHAs by means of DNA
10 shuffling described herein.

- In another embodiment acyltransferase enzymes which specifically esterify the sec-hydroxyl of AHAs by means of active carboxyl transfer from either acyl-coenzyme A or acylated acyl carrier protein (ACP) are incorporated into the reaction pathway. This pathway is depicted in Fig 4, as shown by the coupling of compounds X and XII to yield
15 compound XIII. A preferred embodiment of this pathway, involves recruiting and expressing gene(s) encoding acyl-CoA-dependent acyltransferases, including those which utilize as substrates acetyl-CoA and CoA derivatives of fatty acids, as well as lactoyl-CoA, CoA-thioesters with other AHAs, and CoA derivatives of aromatic, arylalkanoic, branched chain alkanolic carboxylic acids, and alpha-aminoacids. Where carboxylic acids (either in
20 from of free acid, salt or ester), intended for esterification of AHAs, are supplied exogenously, or are co-produced by another co-functioning biotransformation or fermentative pathway in the same host organism, or a different host organism, the invention provides a means for facilitating ester formation by recruiting and co-expressing those acyl-CoA ligases or ACPs which effect *in-vivo* activation of these acids forming suitable
25 substrates for the acyl transferase enzymes that act on the AHAs.

- The invention also provides for another type of biochemical transformation of AHAs to AHA carboxylic esters wherein free AHAs are first converted to their active ester form by means of the enzymatic formation of a derivative with CoA or ACP (Fig. 4, compound XIV). Several alternative acyltransferase enzymes (and genes encoding them)
30 can be recruited for effecting subsequent transformations of compound XIV to esters of different compositions. These preferably include AHA-CoA transferases acting (a) on alcohols (XX) to produce esters (IX), or (b) on molecule of compound XIV or compound XV to produce acyclic homo- and hetero- oligomers (n=2-5) of AHAs. By recruiting an

additional thioesterase enzymes, the activated forms of these oligomeric esters can be converted to free carboxylic oligomers (e.g., XVIII) or to the cyclic substituted glycolides (XVI).

In another preferred embodiment, the formation of an α -hydroxycarboxylic acid ester is catalyzed by an acyl CoA-ligase that is evolved by nucleic acid shuffling. In a preferred embodiment, shuffling of nucleic acids encoding acyl-CoA ligase activities results in an increase in the synthesis of esters. In another preferred embodiment, the esters are selected from structures XIII-XVIII (Fig. 4). The synthesis of these and other esters will generally rely on the provision of a corresponding α -hydroxycarboxylic acid precursor. In a preferred embodiment, the α -hydroxycarboxylic acid precursor is present in an amount sufficient to establish intracellular pools of CoA-activated carboxylic derivatives of α -hydroxycarboxylic acids.

In still another preferred embodiment, the transferase polypeptide is selected from glycosyltransferase and methyltransferase, more preferably methyltransferase and more preferably still a S-adenosylmethionine dependent O-methyltransferase.

5. *Enzymes effecting chiral switch at the level of AHAs.*

Another object of this invention is the effective control of the enantiomeric composition of the compounds prepared by the methods of the invention. For clarity of illustration, the discussion below focuses on AHA esters made by the biotransformation process from alkenes. This focus is intended to be illustrative and not limiting of the scope of this embodiment of the invention.

Means of enantiomeric control, when integrated as part of the multistep biocatalytic pathway, constitutes an important advantage as it allows selective production of either enantiomer of the AHA. The enantiomerically pure AHAs can be used as resolving reagents, chiral synthons, or monomers for polyesters or co-polyesters with lactic acid.

In a preferred embodiment, the AHA is mandelic acid, or an analogue thereof, and the chiral switch is effected by recruiting mandelate a racemase gene.

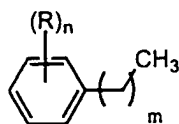
Mandelate racemase catalyzes the interconversion of the R and S enantiomers of mandelic acid and its derivatives. An exemplary mandelate racemase is that of *Pseudomonas putida* (the sequence of the gene can be found in the GenBank database under the locus [PSEMDLABC]). Preferred mandelate racemases are those of the *P. putida* strain ATCC 12633, however, mandelate racemases from any other organism can be used.

Although, in a preferred embodiment, the chiral switch is made at the level of the AHA, this switch can be made with any of the precursors or adducts of the AHA as well. Thus, in yet another preferred embodiment, the AHA is modified by at least one of the ester-forming enzymes discussed herein. Preferred ester forming enzymes are those which specifically, or preferentially, act on one enantiomer of the AHA, thus allowing enantiospecific resolution of the racemate *in-vivo*. The activity of the above racemases provides an enantiomeric equilibrium at the expense of the non-esterified enantiomer. The combined action of the racemase and the AHA esterifying enzymes provides a chiral switch which allows preparation of one desired enantiomer, whether R or S, from AHAs of any enantiomeric composition.

6. Hydroxylation of organic substrates

The monooxygenase polypeptides of the invention are capable of hydroxylating substantially any substrate comprising a terminal methyl, internal methylene or π -bond group. These substrates include, for example, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl and the like. Other appropriate substrates will be apparent to those of skill in the art.

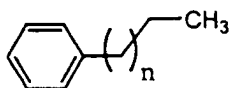
In a preferred embodiment, the substrate has the structure:



wherein, each of the n R groups is a member selected from the group consisting of H, alkyl groups and substituted alkyl groups; m is a number from 0 to 10, inclusive; and n is a number from 0 to 5, inclusive.

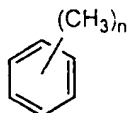
In another preferred embodiment, the substrate includes benzene substituted with a member selected from the group of straight-chain alkyl groups branched-chain alkyl groups and combinations thereof. The substituent is more preferably, a member selected from C_1 - C_6 straight-chain, C_1 - C_6 branched-chain alkyl and combinations thereof, and even more preferably, ethyl, *n*-propyl, *i*-propyl, *t*-butyl and combinations thereof.

In another preferred embodiment, the substrate has the structure:



wherein, n is a number between 0 and 9, inclusive.

In yet another preferred embodiment, the substrate has the structure:



wherein, n is an integer from 1 to 6.

5 Presently preferred products of these oxidation reactions include benzyl alcohol, substituted benzyl alcohol, 2-phenylethanol, substituted 2-phenylethanol, 3-phenylpropanol, substituted 3-phenylpropanol and their derivatives.

10 In a still further preferred embodiment, the substrate includes a member selected from 3,4-dihydrocoumarin and 3,4-dihydrocoumarin residues and the poly peptide converts a methylene group of the substrate to $—CH(OH)—$.

In yet another preferred embodiment, the substrate is 3,4-dihydrocoumarin and the polypeptide converts the substrate to 4-hydroxy-4-dihydrocoumarin.

7. Preparation of hydroxylated aromatic carboxylic acids

15 Hydroxylated aromatic carboxylic acids have many diverse uses, including as antimicrobial additives, UV protectants (e.g. esters of p-hydroxybenzoic acid, parabens), pharmaceutical compositions (e.g., esters of salicylic acid, coumarins and 3,4-dihydroxycoumarin).

20 Thus, in another preferred embodiment, the present invention provides a method for preparing hydroxylated aromatic carboxylic acids. The method includes contacting a substrate comprising an aryl carboxylic acid with a dioxygenase polypeptide of the invention. The polypeptide is preferably expressed by an organism of the invention.

a. Carboxylic acid substrates

25 The carboxylic acids used as substrates in the present invention can be obtained from commercial sources, or they can be prepared by methods known in the art. In a preferred embodiment, the carboxylic acids are prepared by contacting a substrate comprising an aryl alkyl group with an oxygenase polypeptide to produce the corresponding aryl alkyl alcohol. The alcohol is subsequently acted upon by a dehydrogenase polypeptide

to produce the desired carboxylic acid. Alternatively, the alcohol can be converted to COOH by chemical means.

For clarity of illustration, the discussion herein focuses on the oxidation of arylmethyl groups to carboxylic acids. This focus is intended to be illustrative and not limiting.

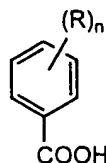
(i). Alkyl group monooxygenation

The first step in the biotransformation processes for conversion of alkylaryl compounds, such as toluene and isomeric xylenes includes the selective oxidation of at least one methyl group present in the aromatic substrate to the corresponding carboxylic acid (e.g., benzoic, toluic acids). In an exemplary embodiment, the substrate is a *p*- or a *m*-xylenes and preferably, only one of the methyl groups is oxidized.

Following the oxygenation step, the resulting alcohol is dehydrogenated, generally by the action of a dehydrogenase polypeptide to produce the desired carboxylic acid.

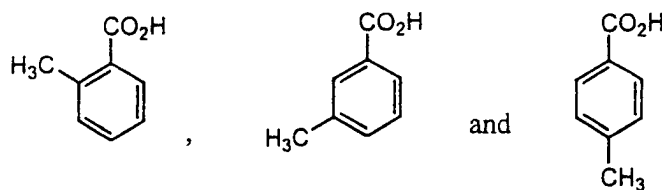
The invention provides for polypeptides that selectively oxidize only one alkyl group of an arene bearing two or more alkyl substituents. In an exemplary embodiment, xylene is converted to a monocarboxylic acid. Alternatively, the invention provides polypeptides that are capable of oxidizing more than one alkyl substituent of a species substituted with two or more alkyl groups. This is in contrast to certain polypeptides of the invention are capable of oxidizing both of the methyl substituents of a xylene to produce the corresponding benzenedimethanol (4a).

In a preferred embodiment, the monooxygenation/dehydrogenation pathway produces a carboxylic acid having the structure:



wherein each of the *n* R groups is independently selected from H, alkyl and substituted alkyl groups; and *n* is a number from 1 to 5, inclusive, more preferably R is methyl, and more preferably still, *n* is a number from 1 to 3, inclusive.

In a still further preferred embodiment, the carboxylic acid is selected from:



Many enzymes for effecting these reactions are well known in the art, and are suitable for use in the construction of useful polypeptides and host strains. To achieve the initial oxidation of the methyl groups, certain enzymes are presently preferred, including

5 non-heme multicomponent monooxygenases of toluene and xylenes, and *p*-cymene, as well as certain arene dioxygenases which act on these substrate in a monooxygenase mode. The latter are exemplified by naphthalene dioxygenase, 2-nitrotoluene 2,3-dioxygenase and 2,4-dinitrotoluene 4,5-dioxygenase. These dioxygenases do not oxidize the aromatic ring of methylbenzenes, but are capable of oxidizing methyl groups of a variety of

10 aromatic compounds in a monooxygenase mode (Selifonov, *et al.*, *Appl. Environ. Microbiol.* 62(2):507-514 (1996); Lee *et al.*, *Appl. Environ. Microbiol.* 62(9):3101-3106 (1996); Parales, *et al.*, *J. Bacteriol.* 180(5):1194-1199 (1998); Suen *et al.*, *J. Bacteriol.* 178(16):4926-4934 (1996). As with the other polypeptide activities discussed herein, the ability of a dioxygenase to act as a monooxygenase is a property that can be optimized by shuffling the

15 nucleic acids encoding these dioxygenases.

The following list provides examples of polynucleotides that encode dioxygenases acting as monooxygenases and which are suitable for use in the methods of the invention. The loci are identified by GenBank ID and encode complete or partial protein components of the arene dioxygenases. Suitable loci include:

20 [AB004059], [AF010471], [AF036940], [AF053735], [AF053736], [AF079317], [AF004283], [AF004284], [PSENAPDOXA], [PSENAPDOXB], [PSENDOABC], [PSEORF1], [PSU49496] naphthalene-1,2-dioxygenase; [BSU62430] 2,4-dinitrotoluene dioxygenase; [PSU49504] 2-nitrotoluene dioxygenase.

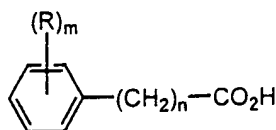
25 The polypeptide that catalyzes the monooxygenation can be a naturally occurring polypeptide, or it can have one or more properties that are improved relative to an analogous naturally occurring polypeptide. In a preferred embodiment, the polypeptides are expressed by one or more host organisms. Moreover, the polypeptide that catalyzes the monooxygenation can be co-expressed by the same host expressing a polypeptide used for

further structural elaboration of the oxidation substrate or product (*e.g.*, a dioxygenase polypeptide that oxidizes the π -bond). Alternatively, the mono- and di-oxygenase polypeptides can be expressed in different hosts.

5 (ii). *Oxidation of alkylarenes having alkyl groups with $\geq C_2$*

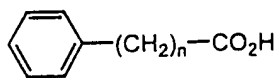
While much of the discussion above highlighting pathway and organism construction for oxidation of methylbenzenes is directly applicable to the set of processes dealing with alkyl benzenes bearing other alkyl groups.

Thus, in a preferred embodiment, at least one alkyl group of the alkylarene
10 has at least two carbon atoms. Preferred species produced in the monooxygenation step (and any subsequent structural elaboration) have the structure:



wherein each of the m R groups is selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl and substituted heterocyclyl; m is a
15 number from 0 to 5, inclusive; and n is a number from 1 to 10, inclusive. Preferred aryl groups are those substituted on the aryl group with at least one methyl moiety.

In another preferred embodiment, the compound has the structure:



wherein n is a number from 1 to 6, inclusive.

20 Generally, oxidation of C_2 alkyl groups is best accomplished by expressing a suitable cytochrome P450 type enzyme system. The enzymes of this class are ubiquitous in nature, and they can be found in a variety of organisms. For example, *n*-propylbenzene is known to undergo α -oxidation in strains of *Pseudomonas desmolytica* S449B1 and *Pseudomonas convexa* S107B1 (Jigami *et al.*, *Appl. Environ. Microbiol.* **38**(5):783-788
25 (1979)) which can utilize this hydrocarbon in either of two alternative oxidation pathways.

Similarly, well known in the art, alkane monooxygenases of bacterial origin, or cytochromes P450 for camphor oxidation, whether wild-type or mutant, can be recruited for the purpose of introducing the oxygen at the terminal methyl group of alkylarenes (Lee *et al.*, *Biochem. Biophys. Res. Commun.* **218**(1):17-21 (1996); van Beilen *et al.*, *Mol.*

Microbiol. 6(21):3121-3136 (1992); Kok *et al.*, *J. Biol. Chem.* 264(10):5435-5441 (1989); Kok *et al.*, *J. Biol. Chem.* 264(10):5442-5451 (1989); Loida *et al.*, *Protein Eng.* 6(2):207-212 (1993).

5 (iii) *Oxygenation of arenes with exocyclic π -bonds*

In another preferred embodiment, the starting material for the carboxylic acid is an arene bearing an exocyclic π -bond. This class of compounds is exemplified by styrene. Other analogous species are set forth in Fig. 3.

The conversion of the exocyclic π -bond is best accomplished by recruiting a
10 cluster of bacterial styrene oxidation genes well known in the art (Marconi *et al.*, *Appl. Environ. Microbiol.* 62(1):121-127 (1996); Beltrametti *et al.*, *Appl. Environ. Microbiol.* 63(6):2232-2239 (1997); O'Connor *et al.*, *Appl. Environ. Microbiol.* 63(11):4287-4291 (1997); Velasco *et al.*, *J. Bacteriol.* 180(5):1063-1071 (1998); Itoh, *et al.*, *Biosci. Biotechnol. Biochem.* 60(11):1826-1830 (1996). Alternatively, the styrene epoxidation step
15 can be accomplished by using monooxygenases active towards methyl substituted aromatic compounds, such as toluene or xylenes (Wubbolts, *et al.*, *Enzyme Microb. Technol.* 16(7):608-615 (1994).

(iv) *Dehydrogenation*

20 To produce the desired carboxylic acid, the alcohol from (i-iii), above, is preferably treated with a dehydrogenase polypeptide. The dehydrogenase enzymes can be endogenous to a host that expresses one or more of the oxygenase polypeptides, or it can exhibit properties that are improved relative to an endogenously expressed dehydrogenase.

The polypeptide that catalyzes the dehydrogenation can be a naturally
25 occurring polypeptide, or it can have one or more properties that are improved relative to an analogous naturally occurring polypeptide. In a preferred embodiment, the polypeptides are expressed by one or more host organisms. Moreover, the polypeptide that catalyzes the dehydrogenation can be co-expressed by the same host expressing one or more of the dioxygenase polypeptide. Alternatively, the dehydrogenase and oxygenase polypeptides can
30 be expressed in different hosts.

In yet another preferred embodiment, the invention provides a method for altering or controlling the regiospecificity of the dehydrogenation reaction of a vicinal diol. This method "blocks" one of the vicinal diol hydroxyl groups by forming an ester, for

example. The method includes contacting the vicinal diol with a polypeptide, preferably expressed by a host organism, having an activity selected from ligase, transferase and combinations thereof, thereby forming a α -hydroxycarboxylic acid adduct. As with the other polypeptides discussed above, this polypeptide can be expressed by the same host cell that expresses other polypeptides of the reaction cascade. Moreover, this polypeptide can be a naturally occurring polypeptide, or it can be improved using the method of the invention.

b. Monooxygenation of aromatic π -bonds

In the synthesis of hydroxyaryl carboxylic acids using the methods of the invention, once the carboxylic acid moiety is in place, the molecule is submitted to an arene monooxygenation cycle (Fig. 1). The monooxygenation of the aromatic ring is preferably accomplished by recruiting one or more monooxygenase genes, preferably of bacterial origin. Exemplary monooxygenase genes are disclosed herein. The method of the invention can be practiced using essentially any type of aromatic ring system. Exemplary aromatic systems include, benzenoid and fused benzenoid ring systems (*e.g.*, benzene, naphthalene, pyrene, benzopyran, benzofuran, *etc.*) and heteroaryl systems (pyridine pyrrole, furan, *etc.*). In a preferred embodiment, the substrate includes a benzenoid hydrocarbon.

Similar to the embodiments discussed above, in this embodiment, the polypeptide that catalyzes the monooxygenation can be coexpressed with one or more polypeptides used in a synthetic pathway. For example, the monooxygenase, dehydrogenase and transferase polypeptides can all be coexpressed in a single host. Other functional combinations of coexpression will be apparent to those of skill in the art.

3. *Conversion of hydroxyls and/or acids to esters*

In another preferred embodiment, there is provided a method for converting carboxylic acid and hydroxyl groups to adducts such as esters and ethers. Useful polypeptides include ligases and transferases (*see*, Fig. 4). For the purposes of the discussion below, these polypeptides are referred to as "adduct-forming" polypeptides.

The adduct-forming polypeptides are useful for enhancing the production of biotransformation products. These polypeptides, which convert a diol, for example, to a monoacyl or monoglycosyl derivative, can enhance control over the regioselectivity of subsequent reactions (*e.g.*, chemical dehydration). For example, the regioselectivity of chemical dehydration in certain cases can be controlled by converting the compounds to

their diacyl derivatives by means of chemical reaction, and then selectively removing one of the acyl groups using an polypeptide of the invention. Alternatively, one can control the regioselectivity of the dehydration by using an esterase or a trans-acylase polypeptide to convert the compounds to monoacyl derivatives in the presence of an excess of another
5 carboxylic acid ester, in an essentially organic medium. In addition, acylation of diols, for example, to obtain monocarboxylic esters provides advantages for efficient recovery of such esters by means of organic solvent extraction, including by extraction with organic solvents which may be used in an immiscible biphasic organic-aqueous biotransformation with whole cells, whether in a batch or in a continuous mode.

10 An adduct-forming polypeptides can be expressed by the same host cell that expresses the monooxygenase, dehydrogenase, racemase, *etc.*, or it can be expressed by a different host cell. Moreover, an adduct-forming polypeptide can be a naturally occurring polypeptide, or it can be improved by the method of the invention.

When the adduct-forming polypeptide is an improved polypeptide, in
15 presently preferred embodiments, the polypeptides can, for example, demonstrate increased efficiency in the formation of the monoacyl- or monoglycosyl- derivatives of a desired compound (*e.g.*, a glycol, carboxylic acid, *etc.*). Other improved adduct-forming polypeptides include transferases and ligases that can selectively modify only one of the hydroxyl groups of a diol, thus providing a means for control of regioselectivity of
20 dehydration of such derivatives to either of two possible isomeric α -hydroxycarboxylic acid compounds.

4. *Conversion of fatty acids to hydroxy acids*

In another preferred embodiment, there is provided a method for converting
25 fatty (preferably, alkanolic, $n=3-20$) acids to hydroxy acids. Monooxygenases are well known to those skilled in the art to perform the oxidation of remote carbons in a fatty acid. Improved polypeptides will have selectivity for the oxidation of any position in the chain. These hydroxyacids can then be used as substrates for polymer formation.

30 D. **Antioxidant and Impurity Modification and Detoxification**

In another embodiment, the invention provides a means for degrading or modifying organic materials which leads to their detoxification. Exemplary compounds include stabilizing agents, antioxidizing agents, environmental pollutants and the like. This

method is applicable to substantially any compound that can be detoxified by, for example, oxidation, either with or without additional structural elaboration. For clarity of illustration, the discussion below focuses on the detoxification of agents commonly found in organic solvents and in π -bonded compounds of use in the present invention.

5 Many commercially available compounds (*e.g.*, alkylbenzenes, alkenes, *etc.*) are stabilized with small amounts of antioxidants such as 4-*tert*-butylcatechol or alkylphenols (*e.g.* BHT) to prevent polymerization during storage and transportation. While the amount of these compounds is usually relatively small (10-15 ppm), they can inhibit biocatalyst performance as they accumulate in aqueous fermentation medium during
10 prolonged incubations required to obtain satisfactory endproduct concentrations.

Several types of enzymes for modifying the phenolic stabilizing compounds can be used to alleviate any negative effects of these compounds on the whole cell biocatalyst performance. Their genes can be introduced in the same host organism used to produce endproducts or intermediate of relevance to his invention. Alternatively, they can
15 be incorporated into a separate host organism. This obviates the need for additional steps in the process which may be required in order to remove these stabilizers. Optimization of one or several of these enzymes for the efficient removal of these stabilizing compounds is a target for DNA shuffling.

Exemplary enzymes for modifying phenolic and diphenolic stabilizers
20 include, but not limited to, acyltransferase, methyltransferase, glycosyltransferase, lactase and peroxidase. In addition to these enzymes, catecholic stabilizers also can be modified to innocuous products by catechol dioxygenases effecting *meta*- or *ortho*-ring cleavage. Many of these enzymes show a significant breadth of activity towards compounds related to phenolic stabilizers. Thus, DNA shuffling can be applied to optimize enzyme parameters
25 such as:

- a) increased turnover with particular phenolic stabilizer,
- b) increased functional expression, by obviating the requirements for certain post-translational modifications of those enzymes which require such modifications (*e.g.* glycosylation of peroxidases and lactases); and
- 30 c) alleviation of inhibition of these enzymes by high concentration of co-occurring feedstock compounds and intermediates and endproducts of the biocatalytic process.

E. Analytical Methodology

A number of analytical techniques are useful in practicing the present invention. These analytical techniques are used to measure the extent of conversion of a particular substrate to product. These techniques are also used to analyze the regioselectivity and/or the enantiomeric selectivity of a particular reaction catalyzed by a polypeptide of the invention. Moreover, these techniques are employed to assess the effect of nucleic acid shuffling experiments on the efficiency and selectivity of the polypeptides produced following the shuffling. The discussion below focuses on those aspects and embodiments of the invention in which an olefin precursor is oxidized by a monooxygenase. The analytical techniques discussed in this context are generally of broad applicability to other aspects and embodiments of the invention. This is particularly true of the spectroscopic and chromatographic methods discussed below. Thus, in the interest of brevity, the following discussion focuses on analyzing the products of the oxidation of an olefin, but the utility of the methods discussed is not limited to this embodiment.

1. Selecting for Monooxygenase activity

Monooxygenase activity can be monitored by HPLC, gas chromatography and mass spectroscopy, as well as a variety of other analytical methods available to one of skill. The consumption of molecular oxygen by the monooxygenase can be measured using an oxygen sensing system, such as an electrode. Incorporation of ^{18}O from radio-labeled molecular oxygen can be monitored directly by mass shift by MS methods and by an appropriate radioisotope detector with HPLC and GC devices. For example, epoxidation of 1-hexadecene to 1,2-epoxyhexadecene can be monitored by ^{18}O incorporation either in intact whole cell or lysate. This has been used, for example by Bruyn et al with *Candida lipolytica*.

In addition, epoxide formation can be indirectly measured by various reactive colorimetric reactions. When H_2O_2 is used as the oxidant, disappearance of peroxide over time can be monitored directly either potentiometrically or colorimetrically using a number of commercially available peroxide reactive dyes.

In a high-throughput modality, the method of choice is high-throughput MS, or MS with an electron spray-based detection method. In addition, selection protocols in which the organism uses a given alkane, alkene or epoxide as a sole carbon source can be

used. In some systems this will be most readily accomplished by combining the alkene oxidizing polypeptide with an epoxide hydrolase to generate a metabolizable alcohol.

2. *Automation for Strain Improvement*

5 One key to strain improvement is having an assay that can be dependably used to identify a few mutants out of thousands that have potentially subtle increases in product yield. The limiting factor in many assay formats is the uniformity of library cell (or viral) growth. This variation is the source of baseline variability in subsequent assays. Inoculum size and culture environment (temperature/humidity) are sources of cell growth
10 variation. Automation of all aspects of establishing initial cultures and state-of-the-art temperature and humidity controlled incubators are useful in reducing variability. In one aspect, library members, *e.g.*, cells, viral plaques, spores or the like, are separated on solid media to produce individual colonies (or plaques). Using an automated colony picker (*e.g.*, the Q-bot, Genetix, U.K.), colonies are identified, picked, and 10,000 different mutants
15 inoculated into 96 well microtitre dishes containing two 3 mm glass balls/well. The Q-bot does not pick an entire colony but rather inserts a pin through the center of the colony and exits with a small sampling of cells, (or mycelia) and spores (or viruses in plaque applications). The time the pin is in the colony, the number of dips to inoculate the culture medium, and the time the pin is in that medium each effect inoculum size, and each can be
20 controlled and optimized. The uniform process of the Q-bot decreases human handling error and increases the rate of establishing cultures (roughly 10,000/4 hours). These cultures are then shaken in a temperature and humidity controlled incubator. The glass balls in the microtiter plates act to promote uniform aeration of cells and the dispersal of mycelial fragments similar to the blades of a fermenter.

25

a. Prescreen

The ability to detect a subtle increase in the performance of a shuffled library member over that of a parent strain relies on the sensitivity of the assay. The chance of finding the organisms having an improvement is increased by the number of individual
30 mutants that can be screened by the assay. To increase the chances of identifying a pool of sufficient size, a prescreen that increases the number of mutants processed by 10-fold can be used. The goal of the primary screen will be to quickly identify mutants having equal or

better product titres than the parent strain(s) and to move only these mutants forward to liquid cell culture for subsequent analysis.

In one preferred embodiment, the prescreen for P450 activity is a method for measuring functional heme incorporation. Active P450 monooxygenases have an absorbance at around 450 nm in the presence of carbon monoxide in a reducing environment. Thus expression of the P450 library on an agar plate is followed by the addition of a reducing solution, such as dithionite in water. This solution is then removed and the plate is placed in a CO atmosphere. Colonies with increased absorbance at 450 nm are picked as active cytochrome P450 enzymes. This screening process is general for all P450 monooxygenases.

3. Selection for Redox Partners

One target for the application of gene shuffling technologies is to evolve monooxygenases to use cheaper, more practical redox partners. However, the complexities of managing redox equivalents can be circumvented, in many cases, by using peroxides (such as hydrogen peroxide) as co-substrates. For example, a monooxygenase capable of oxidizing 1-octene to 1,2-epoxyoctane does so in a non-NAD(P)H-dependent manner when H₂O₂ is added to the reaction mix. For peroxidases and chlorperoxidases this peroxide-dependent, NAD(P)H-free oxidative chemistry is the norm. Peroxide-mediated oxidations, however, often result in the rapid inactivation of catalytic activity by a variety of partially understood mechanisms enzymes (*see*, CYTOCHROME P450: STRUCTURE, MECHANISM, AND BIOCHEMISTRY [2nd edition], P.R. Ortiz de Montellano, editor, New York: Plenum Press, chapter 9; and Meunier, B. *Chem. Rev.* 92:1411-1456 (1992)). Enhancing the stability of P450 enzymes in the presence of peroxides and increasing the overall turnover rates of these enzymes with basic industrial raw materials is a feature of the invention.

Gene shuffling offers a means of generating new peroxidase and oxygenase polypeptides with altered selectivity, activity or stability. Whereas peroxides are often prohibitively expensive for use as oxidants for industrial chemistry, biological systems offer the potential to generate and use peroxides *in situ* without isolation of the reactive intermediates. The concepts disclosed here include the coevolution of a hydrogen peroxide-generating system (such as glucose, galactose or alcohol oxidases) with a monooxygenase polypeptide capable of using the peroxide generated to synthesize an oxidized coproduct. In

this context, peroxides can be commercially feasible oxidizing agents for even low-value, high-volume commodity chemicals.

4. *Screening for improved monooxygenase activity.*

In each of the aspects and embodiments discussed below, the concept of screening the library of recombinant polypeptides to enable the selection of improved members of the library is set forth. Although it will be apparent to those of skill in the art that many screening methodologies can be used in conjunction with the present invention, the invention provides a screening process comprising:

- (a) introducing the library of recombinant polynucleotides into a population of test microorganisms such that the recombinant polynucleotides are expressed;
- (b) placing the organisms in a medium comprising at least one substrate;
- and
- (c) and identifying those organisms exhibiting an improved property compared to microorganisms without the recombinant polynucleotide.

a. *Oxidation of olefins*

Depending on the specific outcome desired from a particular course of DNA shuffling of nucleic acids encoding oxygenases for biocatalytic oxidation of olefins, the invention provides several methods for detecting and measuring catalytic properties encoded by the recombinant polynucleotides. These are exemplified by the following methods.

For the purpose of the optimization of individual reactions and whole pathways for production of α -hydroxycarboxylic acids, their derivatives, analogues and precursor compounds described in this invention can be monitored by virtually any analytic technique known in the art. In preferred embodiments, the production of the desired compound is monitored using one or more techniques selected from thin layer chromatography (TLC), high performance liquid chromatography (HPLC), chiral HPLC, mass-spectrometry, mass spectrometry coupled with a chromatographic separation modality, NMR spectroscopy, radioactivity detection from a radioactively labeled compounds (e.g., -olefins, diols, aldehydes, AHAs, etc.), scintillation proximity assays, and by UV-spectroscopy. In a high throughput modality, the preferred methods are selected from one or any combination of these methods.

The methods of the invention are used to improve polypeptides that catalyze the initial oxidation of π -bonded species. Methods using monooxygenase-based pathways are encompassed herein. The oxidation product from the conversion of a substrate comprising a π -bond (*e.g.*, arenes, alkylarenes, alkenes, *etc.*) can be detected by numerous methods well known to those of skill in the art. Certain preferred methods are set forth herein.

In a preferred embodiment, the vicinal diol derived from oxidation of an olefin is quantitated using a radioactively labeled substrate. Although any radioactive isotope commonly used in the art can be incorporated into a substrate, preferred isotopic labels include, for example, ^{14}C and/or ^3H . Differences in the volatility of the olefin substrate and the corresponding diol can be exploited to quantitate the radioactively labeled product. This method can easily be applied to aqueous samples of culture fluids obtained by incubating individual clones of cells expressing libraries of a recombinant polynucleotide obtained using the methods of the invention.

In an exemplary embodiment, cells expressing libraries of recombinant polynucleotides encoding a monooxygenase can be grown in a multiwell dish with a radioactive substrate administered directly to the aqueous medium. After incubation of the cells with the radioactive olefin substrate, any residual unconverted substrate is removed by evaporation, with or without application of vacuum. After removing the unconverted substrate, the culture fluid (or aliquots thereof) is mixed with a suitable scintillation cocktail, and the radioactivity in the samples is quantitatively measured. In a preferred embodiment, selection of the most active clones is based on the amount of radioactivity incorporated into the compounds produced by the organisms expressing the clone.

Alternatively, radioactively labeled substrate can be administered as a vapor phase to colonies growing on a surface of a membrane filter overlaying agar-solidified medium. After incubation, the membrane is removed from the agar surface, and any residual hydrocarbon is evaporated from the membrane. The membrane is autoradiographed, or a scintillation dye is sprayed over the membrane for radioactivity detection. A modification of this assay that is particularly suitable for ^{14}C label detection in and/or around colonies capable of oxidizing π -bonds to the corresponding glycols involves using a porous membrane that has scintillation dye incorporated in the membrane composition by covalent or adsorption means. This assay is termed "scintillation proximity assay on membrane" or "SPA."

In another embodiment of this invention, a variation of SPA is used to selectively quantify the glycol derived from the substrate. This variation involves adding beads for scintillation proximity assay to the samples of culture fluids or extracts obtained by incubation of cells with radiolabeled substrate as described above. Alternatively, the sample
5 can be applied to a membrane. The beads or membrane are functionalized with groups that interact with a glycol.

In a preferred embodiment of this assay, the beads or membranes contain a suitable scintillating dye and their surfaces are modified by chemical groups that interact readily with diols. Such materials can be prepared by known chemical methods from
10 commercially available SPA materials and they can be used to trap free diols directly in the aqueous medium or culture broths obtained by incubation of the microbial cells with the radiolabeled substrates.

In another preferred embodiment, the surface of the beads used in this assay is functionalized with a sufficient amount of a compound that interacts with a glycol, such as
15 compounds containing aryl or alkylboronate (boronic acid). Such beads can be obtained by chemical modification of commercially available SPA beads by reactions known to one skilled in the art. In a preferred embodiment, the reactions used to modify the beads are analogous to those used for the preparation of arylboronate-modified resins for solid-phase extraction or chromatography. After incubation, the beads are washed with a sufficient
20 amount of water or other suitable solvent and subjected to quantitative determination of radioactivity.

One can also determine amounts of glycol produced by oxidation of an π -bond by taking advantage of the reactive nature of the substrate. Samples of culture fluids, or extracts in an appropriate solvent, can be treated with known excess amounts of dilute
25 solutions of, for example, a halogen (Cl_2 , Br_2 , I_2), permanganate salts. The residual excess amount of those reagents, left after reaction with any substrate present, can be measured by chemical methods known in the art for determination of these compounds (*see, for example, VOGEL'S PRACTICAL ORGANIC CHEMISTRY 5th Ed., Furniss et al., Eds., Longman Scientific and Technical, Essex, 1989*).

30 Mass spectrometry can also be used to determine the amount of a vicinal glycol formed due to species encoded by the libraries of shuffled oxygenase genes. Mass spectrometric methods allow ion peaks to be detected. The ion peaks derived from the

vicinal glycol can be readily distinguished from peaks derived from olefin substrates. In a preferred embodiment, coordination ion spray or electrospray mass spectrometry is utilized.

In another preferred embodiment, a compound that interacts with a component of the mixture, preferably the glycol, is utilized to enhance the sensitivity and selectivity of the method. In a presently preferred embodiment, the sample analyzed contains excess arylboronic or alkylboronic acid. Preferred boronic acids are those containing at least one nitrogen atom and include, but are not limited to, dansylaminophenylboronic acid, aminophenylboronic acid, pyridylboronic acid.

The ions detected in the mass spectrum derive from cyclic boronate ester derivatives of the glycols with a boronic acid. The samples are preferably analyzed in non-acidic and non-basic organic solvent or aqueous phase, substantially free of alcohols and other glycols. Other appropriate analytical conditions will be apparent to those of skill in the art.

Another preferred method for quantitating the glycols uses periodic acid or its salts, preferably the sodium salts, to cleave the vicinal glycols to the corresponding aldehydes. In a preferred embodiment, vicinal diols other than the analyte (*e.g.*, carbohydrates) are excluded from the aqueous or organic solvent samples. This is easily attained by using non-carbohydrate carbon sources to grow the microbial cells, and/or by removal of the cells from the media by centrifugation or filtration prior to contacting of the sample with periodate reagent. The periodate reagent can be used in solution, or preferably, immobilized on a solid phase (*e.g.* anion exchange resin). After reacting the glycol with an excess of periodate ion, the amount of free aldehyde groups can be measured by a variety of assays known in the art. In a preferred method, the aldehydes are quantitated by a method based on the formation of a colored hydrazone derivative. Alternatively, when using radioactively labeled olefins for biotransformation, the free aldehydes obtained by this method can be trapped by aldehyde reactive groups (*e.g.*, free amines) on the surface of an appropriately modified SPA beads or membranes.

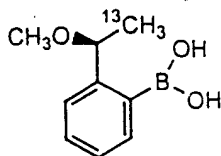
(ii). *Methods for detecting alternative regioselectivity of oxidation of species with multiple π -bonds*

In one embodiment, the substrate includes more than one π -bond (*e.g.*, styrene, butadiene, *etc.*). In a preferred embodiment, one of the π -bonds undergoes reaction more readily than the other. In this embodiment, it is generally preferred to determine which

of the π -bonds underwent reaction. The preferred method for making this determination is ^1H or ^{13}C NMR, although other methods can be used. Other methods include, for example, chromatography (*e.g.*, TLC, GC, HPLC, *etc.*), UV/vis spectroscopy and IR spectroscopy. In an embodiment wherein the reaction is operating in a high throughput mode, the method of choice is a flow-through ^1H or ^{13}C NMR spectroscopy.

When ^{13}C NMR is used, the substrates are preferably labeled with ^{13}C . π -bonded species can be synthesized by methods known in the art from a ^{13}C enriched material to incorporate one, or any combination of several, labeled carbon atom(s) into the structure of these compounds. The enrichment levels for the labeled positions are preferably at least 5% of ^{13}C , more preferably 50% and more preferably still 95% for any given labeled position. Incorporation of a ^{13}C label provides a number of advantages, such as increasing the NMR signal and decreasing time required for spectral acquisition. Moreover, labeled compounds allow for a quantitative or semi-quantitative interpretation of the composition of a mixture of isomeric oxidation products. Preferably, incubations with ^{13}C labeled olefins are conducted in multi-well plates, and aliquots of culture fluids or their extracts are sampled with an autosampler communicating with the NMR probe. In another preferred embodiment, the reaction components are not chromatographed or otherwise purified prior to obtaining a NMR spectrum.

Determining the absolute configuration and the enantiomeric composition of the glycols formed from π -bonded species, preferably employs a variation of the method described above for determining regioselectivity of dihydroxylation of the olefinic substrates by a monooxygenase using ^1H or ^{13}C NMR. In a preferred embodiment, the substrates are labeled with ^{13}C and ^{13}C NMR, is employed. This method preferably involves the use of a chiral and essentially enantiomerically pure derivatizing reagent such as a substituted arylboronic acid which forms a cyclic boronate derivatives with vicinal glycols, as known in the art (references: Resnick, Gibson, 1997, *cite*). In a preferred embodiment, both the substrates and one or more carbon atoms of the boronic acid is labeled with ^{13}C . Although a broad range of boronic acids are of use in the present invention, a currently preferred boronic acid is shown below:



The absolute configuration of any chiral center of the compounds produced by the methods of the invention can be either R or S. In presently preferred embodiments, the enantiomeric excess of the product is preferably 98% or more. NMR signals of different enantiomers of the reaction products can be distinguished in diastereomeric products using substantially enantiomerically pure boronate compounds as discussed above. Moreover, the relative intensity of the NMR signals arising from corresponding atoms of the diastereomeric products can be used for estimating the enantiomeric composition of the product(s) present in the sample.

(iii). *Methods for detecting alternative regioselectivity of oxidation of alkylarenes*

Useful methods for determining the regioselectivity of the oxidation of alkylarene compounds are substantially similar to those described in section (ii), *supra*.

2. *AHA formation from glycols*

Among methods for specifically measuring the free AHAs produced in the biocatalytic process, those which are particularly preferred are methods using a variation of the scintillation proximity assay described above. These methods preferably use an excess of beads or membranes bearing one or more positively charged functional groups (*e.g.* quaternary or tertiary or primary amines). In preferred embodiments, these beads or membranes act as an anion exchange medium and they selectively trap free AHAs, thereby removing them from aqueous culture broths. In another preferred embodiment, this method employs a radioactively labeled starting material, or subsequent intermediate, (*e.g.*, glycol, epoxide, *etc.*). The radioactively labeled compound interacts with the beads or membrane. Prior to measuring the radioactivity associated with the beads or the membrane, non-specifically adsorbed label is preferably removed by evaporating excess radioactive compound and/or washing with an aqueous solution which does not cause elution of the AHAs from the anion-exchange beads or membrane.

Preferred methods for determining the chirality and absolute configuration of AHAs formed in the described biotransformation process are substantially similar to those

methods employed in making these determinations with respect to the glycols, as discussed above.

3. *Methods for determination of HCAs*

5 In HTP mode, a preferred analytical method is flow-through ^1H or ^{13}C NMR spectroscopy. In the ^{13}C NMR mode, the aromatic substrate for oxidation by a monooxygenase is preferably labeled by the ^{13}C isotope. Alkylaryl compounds or the corresponding arylalkanoic acids are synthesized by methods known in the art from a ^{13}C enriched material to incorporate one, or any combination of several, labeled carbon atom(s) into the structure of these compounds. The enrichment levels for any labeled position are preferably at least 5% of ^{13}C , and more preferably at least 95%. Incorporation of ^{13}C label increases sensitivity of the NMR measurement, decreases time required for acquisition of spectrum per sample, and allows for quantitative or semi-quantitative interpretation of compositions of mixtures of isomeric oxidation products. Preferably, incubations with ^{13}C 10 labeled precursors are conducted in multi-well plates, and aliquots of culture fluids or their extracts are sampled with autosampler connected to the solvent line passing through NMR probe without any column separation.

For determining absolute configuration and enantiomeric composition of the HCAs, a variation of the methods described above for determining reaction regioselectivity 20 by ^1H or ^{13}C NMR is used. In conjunction with the preferred use of ^{13}C labeled substrates, ^{13}C NMR is preferably employed.

The absolute configuration of any chiral center may be either R or S. In a preferred embodiment, the enantiomeric excess is 98% or more. NMR signals of different enantiomers of HCAs can be distinguished in diastereomeric products using known methods, 25 such as NMR in conjunction with lanthanide shift reagents -- or after derivatization with Mosher's esters. Alternatively the enantiomeric excess can be determined by chiral GC.

In another preferred embodiment, a variation of the SPA method is used. In this version, a solid support, such as beads or a membrane containing a suitable scintillation dye is used. The solid support is modified with positively charged groups such that it acts 30 like an anion-exchange material. These materials can be prepared from commercially available SPA materials and they can be used to trap free acids directly in the aqueous

medium or culture broths obtained by incubation of the host cells with a radiolabeled alkylarene.

4. *Methods for determination of esters*

5 In the interest of brevity, the following discussion focuses on the determination of esters of AHAs. One of skill will appreciate that the same, or similar, methods can be used to determine esters of other compounds formed using the methods of the invention.

10 Both spectroscopic and non-spectroscopic methods can be used to quantitate the extent of ester synthesis and to characterize the esters. The preferred non-spectroscopic method for assaying AHA methyl ester formation catalyzed by methyl transferases is based on use of a radioactively labeled precursors to AHA methyl esters. ^{14}C or ^3H methyl labeled SAM (or its *in-vivo* precursor, methionine) can be used as a probe. In another preferred embodiment, the labeled substrate is the free α -hydroxycarboxylic acid itself.

15 Using the methods of the invention, methyltransferases that are selective for a particular AHA enantiomer can be selected and further improved by iterative cycles of DNA shuffling and this assay. The selectivity of the methyltransferases of the invention towards a particular enantiomeric configuration of an AHA is preferably measured using samples of the α -hydroxycarboxylic acids that are substantially enantiomerically pure. Host cells
20 employed in this biocatalytic cycle will preferably lack AHA racemase activity (*e.g.* mandelate racemase). In another preferred embodiment, both AHA enantiomers have a different radioactive label, *e.g.* one enantiomer is labeled with ^{14}C , and another with ^3H (at one or more H positions which do not readily exchange with water). Measurement of the radioactivity incorporated into the product is performed using a radioactivity detector that
25 allows for the selective measurement of at least two different isotopes. This variation allows the evaluation of the enantioselectivity of a methyltransferases in a single sample.

The radioactivity associated with methyl esters of AHAs is preferably measured in samples which are obtained by selective extraction or partitioning of the methyl esters from neutral or moderately basic (pH about 6-10) aqueous culture samples. These
30 samples can contain varying amounts of free, labeled AHA, of AHA salts and other non-labeled organic compounds. The samples are preferably obtained by incubating individual clones expressing methyltransferase libraries with the labeled AHAs. The incubation medium is subsequently extracted by adding a defined amount of a preferably water-

immiscible organic solvent, or by contacting the broth with a extraction medium (*e.g.* XAD-1180, or similar beads, or membrane).

In those embodiments employing an extraction medium, following its removal from contact with the broth, the extraction media is preferably washed to remove adventitiously bound compounds. Preferred wash solutions are aqueous that do not elute the AHA methyl esters from the extraction medium, but which remove other molecules adsorbed onto the medium. The radioactivity of the extracted material is then measured by methods well known in the art. In embodiments using beads or a membrane an appropriate scintillating dye is preferably used for detecting the radioactivity.

Substantially similar methods can also be employed for detecting other neutral esters of AHAs, such as those exemplified by glycolides (*e.g.*, XVI, Fig. 13) and esters of type XX. Thus the same approach is useful for assaying and characterizing the ester forming activity of polypeptides represented by libraries of acyl-transferases, or by a combination of AHA-CoA: alcohol acyltransferases and AHA-CoA ligases. Variations on this method can include the use of a radioactively labeled alcohol (*e.g.*, XIX) or any of its *in-vivo* metabolic precursor.

In another preferred embodiment, the method for detecting polypeptide activity leading to the formation of neutral AHA esters employs UV or fluorescence spectroscopy. This method is applicable to those embodiments in which the transferase activity yields products exhibiting distinct UV and/or fluorescent characteristics. Exemplary compounds include, for example, substituted or non-substituted esters of aromatic carboxylic acids (*e.g.*, mandelic acid). In preferred embodiments of this method, a solvent or solid-phase extraction under neutral or moderately basic conditions (pH about 6-12) is performed on the cell culture medium. Compounds thus isolated are detected by measurement of their UV absorption or fluorescence. These spectral parameters are evaluated to determine relative amounts and identities of the products formed by the transferase reactions.

a. Screening for improved transferase activity

The screening of the transferase libraries, obtained by DNA shuffling or other methods as described above, is done most easily in bacterial or yeast systems by one or more of the screening methods described below.

(i). *Methods for detecting increased activity of transferase reactions*

The methods for detection of increased formation of monoacyl- and monoglycosyl-derivatives of, for example, glycols and α -hydroxycarboxylic acids include methods in which physical differences between the substrates, the *cis*-diols and the derivatives arising from the transferase-catalyzed reactions are measured. Preferred methods include HPLC and mass-spectrometry. In a high throughput modality, a method of choice is mass-spectrometry, preferably, coordination ion and/or electrospray mass-spectrometry.

For acyl transferases, another presently preferred method uses a labeled acyl-donor precursor, *e.g.* labeled carboxylic acid or its derivative, administered to the cells that express libraries of shuffled genes encoding acyl ligases and/or acyl transferases, *e.g.*, acyl-CoA ligases and acyl-CoA transferases. The amount of label in the hydrophobic reaction products is measured after extraction of the labeled derivatives into a suitable organic solvent, or after solid-phase extraction of these compounds by addition of a sufficient amount of hydrophobic porous resin beads (*e.g.*, XAD 1180, XAD-2, -4, -8). In the case of a radiolabeled compound, scintillating dye can be present in the organic solvent, added to the samples, or chemically incorporated in the bead polymer. The latter constitutes a modification of scintillation proximity assay method.

(ii) *Methods for detecting regioselectivity of transferase reactions.*

The methods for detecting regioselectivity of the transferase reactions include HPLC, and in an HTP modality, flow-through NMR spectroscopy. When NMR spectroscopy is used for determining relative amounts of different regiomeric monoacyl or monoglycosyl derivatives of oxidized substrates, the latter are preferably obtained by action of the arene monooxygenases on isotopically (^{13}C and/or ^2H) labeled substrate. Another variation of the NMR technique includes use of isotopically labeled precursors of acyl- or glycosyl- donor intermediates.

5. *Selecting for enhanced organic solvent resistance.*

Selection for recombinant polynucleotides that provide improved organic solvent resistance can be accomplished by introducing the library of recombinant polynucleotides into a population of microorganism cells and subjecting the population to a medium that contains various concentrations of the organic hydrophobic compounds of interest. The medium can contain, for example, carbon, nitrogen and minerals, and

preferably does not otherwise limit growth and viability of the cells in the absence of the solvent, thus ensuring that solvent resistance is essentially the only limiting factor affecting growth of the cells expressing variants of the genes encoding solvent resistance traits.

In other embodiments, one can employ a screening strategy to identify those recombinant polynucleotides that encode polypeptides that confer improved solvent resistance. For example, one can screen based on the *in vivo* expression of a reporter gene, such as those encoding fluorescent proteins (exemplified by the green fluorescent protein, GFP). Preferably, for the purpose of detecting the best solvent resistant genes under essentially stationary growth phase conditions, those reporter genes are used which display their function in a fashion dependent on availability of intracellular reducing pools, such as NADH and NADPH, and essentially unimpaired ribosomal biosynthesis of proteins.

Such genes and can be exemplified by several bacterial luciferase gene clusters (*lux*) which contain not only luciferase components, but also all polypeptides required for *in-vivo* regeneration of the aldehyde substrate for luciferase.

A variety of methods can be used to detect and to pick or to enrich for the clones with the most efficient solvent resistant traits as judged by display of the properties associated with the *in-vivo* reporter genes. These methods include, for example, fluorescence activating cell sorting of liquid cell suspensions (*e.g.*, cells that express GFP) and CCD camera imaging of individual colonies grown on a solid(ified) medium (*e.g.*, for cells that express *lux*).

If additional improvement in solvent resistance is desired, one can carry out a series of cycles of iterative DNA shuffling and selection by growing the cells in the presence of the organic solvent. Concentrations of the solvents used for selective growth conditions are incrementally increased after each round of recursive mode DNA shuffling in order to provide more stringent selective pressure for those organisms expressing solvent resistance genes.

For use in a high throughput screening protocol, the increase in the solvent resistance to a particular compound of interest and relevance to the biocatalytic synthesis of interest can also be directly measured by administering a radioactively labeled compound and determining relative distribution of radioactivity between cell biomass and extracellular medium components, similar to the method described by Ramos *et al.*, *J. Bacteriol.* 180:3323-3329 (1998).

F. Bioreactors

In another aspect, the invention provides a bioreactor system for carrying out biotransformations using the improved polypeptides of the invention. The bioreactor includes: (a) an improved monooxygenase polypeptide of the invention; (b) a redox partner
5 source; (c) oxygen; and (d) a substrate for oxidation.

In a preferred embodiment, the monooxygenase polypeptide is an arene monooxygenase polypeptide.

In another preferred embodiment, the bioreactor further includes another useful polypeptide, such as a transferase, ligase, dehydrogenase and the like. The additional
10 useful polypeptide(s) can be co-expressed by a host cell also expressing the improved monooxygenase or it can be expressed by a host cell that does not express the improved monooxygenase. Moreover, each of the polypeptides incorporated into the reactor can be provided as a constituent of a whole cell preparation, a polypeptide extract or as a substantially pure polypeptide. The cells and/or polypeptides can be in suspension, solution
15 or they can be immobilized on an insoluble matrix, bead or other particle. Additional considerations are discussed below. This discussion is intended as illustrative and not limiting. Other bioreactor formats, conditions, *etc.* will be apparent to those of skill in the art.

General growth conditions for culturing the particular organisms are obtained
20 from depositories and from texts known in the art such as *BERGEY'S MANUAL OF SYSTEMATIC BACTERIOLOGY*, Vol.1, N. R. Krieg, ed., Williams and Wilkins, Baltimore/London (1984).

For clarity of illustration, the discussion below focuses on the preferred conditions for the oxidation of an organic substrate using the polypeptides of the invention.
25 It is understood that this focus is for the purpose of illustration and that similar conditions are applicable to pathways of the invention other than oxidation.

The nutrient medium for the growth of any oxidizing microorganism should contain sources of assimilable carbon and nitrogen, as well as mineral salts. Suitable sources of assimilable carbon and nitrogen include, but are not limited to, complex mixtures, such as
30 those constituted by biological products of diverse origin, for example soy bean flour, cotton seed flour, lentil flour, pea flour, soluble and insoluble vegetable proteins, corn steep liquor, yeast extract, peptones and meat extracts. Additional sources of nitrogen are ammonium salts and nitrates, such as ammonium chloride, ammonium sulfate, sodium nitrate and

potassium nitrate. Generally, the nutrient medium should include, but is not limited to, the following ions: Mg^{2+} , Na^+ , K^+ , Ca^{2+} , NH_4^+ , Cl^- , SO_4^{2-} , PO_4^{2-} and NO_3^- and also ions of the trace elements such as Cu, Fe, Mn, Mo, Zn, Co and Ni. The preferred source of these ions are mineral salts.

5 If these salts and trace elements are not present in sufficient amounts in the complex constituents of the nutrient medium or in the water used it is appropriate to supplement the nutrient medium accordingly.

 The microorganism employed in the process of the invention can be in the form of fermentation broths, whole washed cells, concentrated cell suspensions, polypeptide
10 extracts, and immobilized polypeptides and/or cells. Preferably concentrated cell suspensions, *polypeptide* extracts, and whole washed cells are used with the process of the invention (S. A. White and G. W. Claus, *J. Bacteriology* 150:934-943 (1982)).
 Methods of immobilizing polypeptides and cells are well known in the art and include such techniques as microencapsulation, attachment to alginate beads, cross-linked polyurethane,
15 starch particles, polyacrylamide gels and the use of coacervates, which are aggregates of colloidal droplets. In a presently preferred embodiment, the polypeptide and/or cell is immobilized onto a glass particles having a porous outer surface, such as that described in Dubin, *et al.*, U.S. Patent No. 5,922,531, issued July 13, 1999.

 Concentrated washed cell suspensions may be prepared as follows: the
20 microorganisms are cultured in a suitable nutrient solution, harvested (for example by centrifuging) and suspended in a smaller volume (in salt or buffer solutions, such as physiological sodium chloride solution or aqueous solutions of potassium phosphate, sodium acetate, sodium maleate, magnesium sulfate, or simply in tap water, distilled water or nutrient solutions). The substrate is then added to a cell suspension of this type and the
25 oxidation reaction according to the invention is carried out under the conditions described.

 The conditions for oxidizing a substrate in growing microorganism cultures or fractionated cell extracts are advantageous for carrying out the process according to the invention with concentrated cell suspensions. In particular the temperature range is from about 0 °C. to about 45 °C. and the pH range is from about 2 to about 10. There are no
30 special nutrients necessary in the process of the invention. More importantly, washed or immobilized cells can simply be added to a solution of substrate, without any nutrient medium present.

It is also possible to carry out the process according to the invention with polypeptide extracts or polypeptide extract fractions prepared from cells. The extracts can be crude extracts, such as obtained by conventional digestion of microorganism cells. Methods to break up cells include, but are not limited to, mechanical disruption, physical
5 disruption, chemical disruption, and enzymatic disruption. Such means to break up cells include ultrasonic treatments, passages through French pressure cells, grindings with quartz sand, autolysis, heating, osmotic shock, alkali treatment, detergents, or repeated freezing and thawing.

If the process according to the invention is to be carried out with partially
10 purified polypeptide extract preparations, the methods of protein chemistry, such as ultracentrifuging, precipitation reactions, ion exchange chromatography or adsorption chromatography, gel filtration or electrophoretic methods, can be employed to obtain such preparations. In order to carry out the reaction according to the invention with fractionated cell extracts, it may be necessary to add to the assay system additional reactants such as,
15 physiological or synthetic electron acceptors, like NAD^+ , NADP^+ , methylene blue, dichlorophenolindophenol, tetrazolium salts and the like. When these reactants are used, they can be employed either in equimolar amounts (concentrations which correspond to that of the substrate employed) or in catalytic amounts (concentrations which are markedly below the chosen concentration of substrate). If, when using catalytic amounts, it is to be ensured
20 that the process according to the invention is carried out approximately quantitatively, a system which continuously regenerates the reactant which is present only in a catalytic amount must also be added to the reaction mixture. This system can be, for example, a polypeptide which ensures reoxidation (in the presence of oxygen or other oxidizing agents) of an electron acceptor which is reduced in the course of the reaction according to the
25 invention.

If nutrient media is used with intact microorganisms in a growing culture, nutrient media can be solid, semi-solid or liquid. Aqueous-liquid nutrient media are preferably employed when media is used. Suitable media and suitable conditions for cultivation include known media and known conditions to which substrate can be added.

30 The substrate to be oxidized in the process of the invention can be added to the base nutrient medium either on its own or as a mixture with one or more oxidizable compounds. Additional oxidizable compounds which can be used include polyols, such as sorbitol or glycerol.

If one or more oxidizable compounds are added to the nutrient solution, the substrate to be oxidized can be added either prior to inoculation or at any desired subsequent time (between the early log phase and the late stationary growth phase). In such a case the oxidizing organism is preferably pre-cultured with the oxidizable compounds. The inoculation of the nutrient media is effected by a variety of methods including slanted tube cultures and flask cultures.

Contamination of the reaction solution should be avoided. To avoid contamination, sterilization of the nutrient media, sterilization of the reaction vessels and sterilization of the air required for aeration is preferably undertaken. It is possible to use, for example, steam sterilization or dry sterilization for sterilization of the reaction vessels. The air and the nutrient media can likewise be sterilized by steam or by filtration. Heat sterilization of the reaction solution containing the substrate is also possible.

The process of the invention can be carried out under aerobic conditions using shake flasks or aerated and agitated tanks. Preferably, the process is carried out by the aerobic submersion procedure in tanks, for example in conventional fermentors. It is possible to carry out the process continuously or with batch or fed batch modes, preferably the batch mode.

It is advantageous to ensure that the microorganisms are adequately brought into contact with oxygen and the substrate. This can be effected by several methods including shaking, stirring and aerating.

If foam occurs in an undesired amount during the process, chemical foam control agents, such as liquid fats and oils, oil-in-water emulsions, paraffins, higher alcohols (such as octadecanol), silicone oils, polyoxyethylene compounds and polyoxypropylene compounds, can be added. Foam can also be suppressed or eliminated with the aid of mechanical devices.

G. Kits

Also provided is a kit or system utilizing any one of the selection strategies, materials, components, methods or substrates hereinbefore described. Kits will optionally additionally include instructions for performing methods or assays, packaging materials, one or more containers which contain assay, device or system components, or the like.

In an additional aspect, the present invention provides kits embodying the methods and apparatus herein. Kits of the invention optionally include one or more of the

- following: (1) a shuffled component as described herein; (2) instructions for practicing the methods described herein, and/or for operating the selection procedure herein; (3) one or more monooxygenase assay component; (4) a container for holding monooxygenase nucleic acids or polypeptides, other nucleic acids, transgenic plants, animals, cells, or the like and,
- 5 (5) packaging materials.

In another preferred embodiment, the kit provides a library of improved P-450s, that have been produced by shuffling for improved stability, ease of handling, *etc.* The polypeptides in this library have catalytic activities that are substantially identical to those P-450 found in microsome preparations used to screen drugs and other xenobiotic compounds.

- 10 In a further embodiment, the present invention provides for the use of any component or kit herein, for the practice of any method or assay herein, and/or for the use of any apparatus or kit to practice any assay or method herein.

- In yet another embodiment, the kit of the invention includes one or more improved monooxygenase polypeptides of the invention. In a preferred embodiment, the kit
- 15 includes a library of improved monooxygenase polypeptides.

- It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to included within the spirit and purview of this application and are considered within the scope of the appended claims. All
- 20 publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

WHAT IS CLAIMED IS:

- 1 1. A method for obtaining a polynucleotide that encodes an improved
2 polypeptide comprising monooxygenase activity, wherein said improved polypeptide has at
3 least one property improved over a naturally occurring monooxygenase polypeptide, said
4 method comprising:
 - 5 (a) creating a library of recombinant polynucleotides encoding
6 a recombinant monooxygenase polypeptide; and
 - 7 (b) screening said library to identify a recombinant
8 polynucleotide that encodes an improved recombinant monooxygenase
9 polypeptide that has at least one property improved over said naturally
10 occurring polypeptide.
- 1 2. The method according to claim 1, wherein said creating a library
2 comprises:
 - 3 shuffling a plurality of parental polynucleotides to produce one or
4 more recombinant monooxygenase polynucleotide encoding said improved property.
- 1 3. The method according to claim 1, wherein said monooxygenase
2 activity is a member selected from alkene epoxidation, alkane hydroxylation, aromatic
3 hydroxylation, N-dealkylation of alkylamines, S-dealkylation of reduced thio-organics, O-
4 dealkylation of alkyl ethers, oxidation of aryloxy phenols, conversion of aldehydes to acids,
5 dehydrogenation, decarbonylation, oxidative dehalogenation of haloaromatics and
6 halohydrocarbons, Baeyer-Villiger monooxygenation, modification of cyclosporins,
7 hydroxylation of mevastatin, oxygenation of sulfonylureas and combinations thereof.
- 1 4. The method of claim 2, wherein at least one of said parental
2 polynucleotides encode at least one monooxygenase activity.
- 1 5. The method of claim 2, wherein said parental polynucleotides are
2 homologous.
- 1 6. The method of claim 2, wherein at least one of said parental
2 polynucleotides does not encode a monooxygenase activity.

1 7. The method of claim 2, wherein said parental monooxygenase
2 polynucleotide encodes a polypeptide or polypeptide subsequence selected from a P450
3 oxygenase, a heme-dependent peroxidase, an iron sulfur monooxygenase, a quinone-
4 dependent monooxygenase and combinations thereof.

1 8. The method of claim 2, wherein a member selected from said parental
2 polynucleotides, said one or more recombinant monooxygenase polynucleotide, said
3 identified recombinant monooxygenase polynucleotide and combinations thereof is cloned
4 into an expression vector.

1 9. The method of claim 1, wherein said identified recombinant
2 monooxygenase polynucleotide has an ability to catalyze an enzymatic reaction using a
3 redox partner other than NADPH.

1 10. The method of claim 2, further comprising:
2 creating a library of recombinant peroxide production activity
3 polynucleotides encoding a recombinant hydrogen peroxide production activity;
4 screening said library to identify a recombinant polynucleotide that encodes
5 an improved hydrogen peroxide production activity; and
6 co-expressing one or more of said identified hydrogen peroxide production
7 activity polynucleotides and said identified recombinant monooxygenase polynucleotide in a
8 cell.

1 11. The method of claim 2, further comprising:
2 creating a library of recombinant epoxide hydrolase activity polynucleotides
3 encoding a recombinant epoxide hydrolase activity;
4 screening said library to identify a recombinant polynucleotide that encodes
5 an improved epoxide hydrolase activity; and
6 co-expressing one or more of said identified recombinant epoxide hydrolase
7 activity polynucleotides and said identified recombinant monooxygenase polynucleotide in a
8 cell.

1 12. The method of claim 2, further comprising:
2 creating a library of recombinant dehydrogenase activity polynucleotides
3 encoding a recombinant dehydrogenase activity;

4 screening said library to identify a recombinant polynucleotide that encodes
5 an improved dehydrogenase activity; and
6 co-expressing one or more of said identified recombinant dehydrogenase
7 activity polynucleotides and said identified recombinant monooxygenase polynucleotide in a
8 cell.

1 13. The method of claim 1, further comprising:
2 creating a library of recombinant transferase activity polynucleotides
3 encoding a recombinant transferase activity;
4 screening said library to identify a recombinant polynucleotide that encodes
5 an improved transferase activity; and
6 co-expressing one or more of said identified recombinant transferase activity
7 polynucleotides and said identified recombinant monooxygenase polynucleotide in a cell.

1 14. The method according to claim 13, wherein said transferase
2 polynucleotide is a member selected from acyltransferases, glycosyltransferases, methyl
3 transferases and combinations thereof.

1 15. The method of claim 2, wherein said plurality of parental
2 polynucleotides are shuffled to produce a library of recombinant polynucleotides comprising
3 one or more library member polynucleotide encoding one or more monooxygenase activity,
4 which library is selected for one or more monooxygenase activity selected from alkene
5 epoxidation, alkane hydroxylation, aromatic hydroxylation, N-dealkylation of alkylamines,
6 S-dealkylation of reduced thio-organics, O-dealkylation of alkyl ethers, oxidation of aryloxy
7 phenols, conversion of aldehydes to acids, dehydrogenation, decarbonylation, oxidative
8 dehalogenation of haloaromatics and halohydrocarbons, Baeyer-Villiger monooxygenation,
9 modification of cyclosporins, hydroxylation of mevastatin, conversion of cholesterol to
10 pregnenolone, and oxygenation of sulfonylureas.

1 16. A library of recombinant polynucleotides comprising one or more
2 monooxygenase activity made by said method of claim 1.

1 17. The library of claim 16, wherein said library is a phage display
2 library.

1 18. An improved monooxygenase encoding nucleic acid prepared by the
2 method according to claim 1.

1 19. The method of claim 2, wherein said parental polynucleotides are
2 shuffled in a plurality of cells, which cells are prokaryotes or eukaryotes.

1 20. The method of claim 2, wherein said parental polynucleotides are
2 shuffled in a plurality of cells, which cells are yeast, bacteria, or fungi.

1 21. The method of claim 2, wherein said parental polynucleotides are
2 shuffled in a plurality of cells; said method optionally further comprises one or more
3 members selected from

4 (a) recombining DNA from said plurality of cells that display
5 monooxygenase activity with a library of DNA fragments, at least one of which undergoes
6 recombination with a segment in a cellular DNA present in said cells to produce recombined
7 cells, or recombining DNA between said plurality of cells that display monooxygenase
8 activity to produce cells with modified monooxygenase activity;

9 (b) recombining and screening said recombined or modified cells to produce
10 further recombined cells that have evolved additionally modified monooxygenase activity;
11 and

12 (c) repeating (a) or (b) until said further recombined cells have acquired a
13 desired monooxygenase activity.

1 22. The method of claim 2, wherein said method further comprises:

2 (a) recombining at least one distinct or improved recombinant polynucleotide
3 with a further monooxygenase activity polynucleotide, which further polynucleotide is
4 identical to or different from one or more of said plurality of parental polynucleotides to
5 produce a library of recombinant monooxygenase polynucleotides;

6 (b) screening said library to identify at least one further distinct or improved
7 recombinant monooxygenase polynucleotide that exhibits a further improvement or distinct
8 property compared to said plurality of parental polynucleotides; and, optionally,

9 (c) repeating (a) and (b) until said resulting further distinct or improved
10 recombinant polynucleotide shows an additionally distinct or improved monooxygenase
11 property.

1 23. The method of claim 2, wherein said recombinant monooxygenase
2 polynucleotide is present in one or more bacterial, yeast, or fungal cells and said method
3 comprises:
4 pooling multiple separate monooxygenase polynucleotides;
5 screening said resulting pooled monooxygenase polynucleotides to
6 identify an improved recombinant monooxygenase polynucleotides that exhibits an
7 improved monooxygenase activity compared to a non-recombinant monooxygenase activity
8 polynucleotide; and
9 cloning said improved recombinant nucleic acid.

1 24. The method of claim 23, further comprising transducing said distinct
2 or improved nucleic acid into a prokaryote or eukaryote.

1 25. The method of claim 2, wherein said shuffling of a plurality of
2 parental polynucleotides comprises family gene shuffling.

1 26. The method of claim 2, wherein said shuffling of a plurality of
2 parental nucleic acids comprises individual gene shuffling.

1 27. A selected shuffled monooxygenase nucleic acid made by said method
2 of claim 2.

1 28. A DNA shuffling mixture, comprising: at least three homologous
2 DNAs, each of which is derived from a polynucleotide encoding a member selected from a
3 polypeptide encoding monooxygenase activity, a polypeptide fragment encoding
4 monooxygenase activity and combinations thereof.

1 29. The DNA shuffling mixture of claim 28, wherein said at least three
2 homologous DNAs are present in cell culture or *in vitro*.

1 30. A method for increasing monooxygenase activity in a cell,
2 comprising: performing whole genome shuffling of a plurality of genomic polynucleotides in
3 said cell and selecting for one or more monooxygenase activity.

1 31. The method of claim 30, wherein said genomic nucleic acids are from
2 a species or strain different from said cell.

1 32. The method of claim 30, wherein said cell is of prokaryotic or
2 eukaryotic origin.

1 33. The method of claim 30, wherein said monooxygenase activity to be
2 selected is alkene epoxidation, alkane hydroxylation, aromatic hydroxylation, N-dealkylation
3 of alkylamines, S-dealkylation of reduced thio-organics, O-Dealkylation of alkyl ethers,
4 oxidation of aryloxy phenols, conversion of aldehydes to acids, dehydrogenation,
5 decarbonylation, oxidative dehalogenation of haloaromatics and halohydrocarbons, Baeyer-
6 Villiger monooxygenation, modification of cyclosporins, hydroxylation of mevastatin,
7 conversion of cholesterol to pregnenolone, oxygenation of sulfonylureas and combinations
8 thereof.

1 34. A method for obtaining a polynucleotide encoding an improved
2 polypeptide acting on a substrate comprising a target group selected from an olefin, a
3 terminal methyl group, a methylene group, an aryl group and combinations thereof, wherein
4 said improved polypeptide exhibits one or more improved properties compared to a naturally
5 occurring polypeptide acting on said substrate, said method comprising:
6 creating a library of recombinant polynucleotides that encoding a
7 monooxygenase polypeptide acting on said substrate; and
8 screening said library to identify a recombinant polynucleotide
9 encoding an improved polypeptide that exhibits one or more improved properties compared
10 to a naturally occurring monooxygenase polypeptide.

1 35. The method according to claim 34, wherein said library of recombinant
2 polynucleotides is created by recombining at least a first form and a second form of a nucleic
3 acid, at least one form encoding said naturally occurring polypeptide or a fragment thereof,
4 wherein said first form and said second form differ from each other in two or more
5 nucleotides.

1 36. The method according to claim 35, wherein said first and second forms
2 of said nucleic acid are homologous.

1 37. The method according to claim 35, wherein at least one of said first
2 and second forms of said nucleic acid does not encode a polypeptide having monooxygenase
3 activity.

1 38. A polypeptide encoded by a polynucleotide according to claim 34.

1 39. The polypeptide according to claim 38 wherein said polypeptide has an
2 activity comprising, converting an olefin to an epoxide.

1 40. The polypeptide according to claim 38, wherein said polypeptide has an
2 activity comprising, converting said terminal methyl group to a hydroxymethyl group.

1 41. The polypeptide according to claim 38, wherein said polypeptide has an
2 activity comprising, converting a methylene group to a hydroxymethylene group.

1 42. The polypeptide according to claim 38, wherein said polypeptide has an
2 activity comprising, converting an aryl group to a hydroxyaryl group.

1 43. The polypeptide according to claim 38, wherein said improved property
2 is selected from:
3 improved regiospecificity of said acting on a substrate, wherein said
4 substrate comprises at least two target groups;
5 enhanced production of a desired enantiomeric form of a reaction
6 product;
7 enhanced expression of said polypeptide by a host cell that comprises
8 said recombinant polynucleotide; and
9 enhanced stability of said polypeptide in said presence of an organic
10 solvent.

1 44. A method of oxidizing a substrate comprising a target group selected
2 from an olefin, a terminal methyl group, a methylene group, an aryl group and combinations
3 thereof, said method comprising contacting said substrate with a polypeptide according to
4 claim 38

1 45. The method according to claim 44, wherein said absolute configuration
2 of a product of said monooxygenase is R, S, or a mixture thereof.

1 46. A method for preparing an epoxide group, said method comprising
2 contacting a substrate comprising a carbon-carbon double bond with a polypeptide according
3 to claim 39.

1 47. A method for preparing a hydroxymethyl group, said method
2 comprising contacting a substrate comprising a terminal methyl group with a polypeptide
3 according to claim 40.

1 48. A method for preparing a hydroxymethylene group, said method
2 comprising contacting a substrate comprising a methylene group with a polypeptide
3 according to claim 41.

1 49. A method for preparing a hydroxyaryl group, said method comprising
2 contacting a substrate comprising an aryl group with a polypeptide according to claim 42.

1 50. An organism comprising a recombinant monooxygenase polynucleotide
2 encoding an improved polypeptide that catalyzes a reaction selected from epoxidation of an
3 olefin, hydroxylation of a terminal methyl group, hydroxylation of a methylene group,
4 hydroxylation of an aryl group and combinations thereof wherein said polypeptide exhibits
5 one property improved relative to a corresponding property of a naturally occurring
6 monooxygenase polypeptide.

1 51. The organism according to claim 50, further comprising an improved
2 transferase polypeptide that exhibits one or more improved properties improved relative to a
3 corresponding property of a naturally occurring transferase polypeptide.

1 52. The organism according to claim 51, wherein said transferase is
2 selected from S-adenosylmethionine dependent O-methyltransferase, acyl-CoA transferase
3 and combinations thereof.

1 53. The organism according to claim 50, further comprising an improved
2 ligase peptide that exhibits one or more properties improved relative to a corresponding
3 property of a naturally occurring ligase polypeptide.

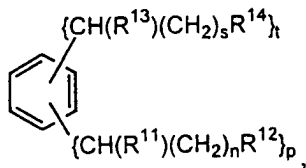
1 54. The organism according to claim 53, wherein said ligase is an acyl
2 CoA ligase.

1 55. The organism according to claim 50, further comprising an improved
2 racemase polypeptide that exhibits one or more properties improved relative to a
3 corresponding property of a naturally occurring racemase polypeptide.

1 56. The organism according to claim 55, wherein said racemase is
2 mandelate racemase.

1 57. The organism according to claim 50, further comprising a
2 dehydrogenase polypeptide that exhibits one or more properties improved relative to a
3 corresponding property in a naturally occurring dehydrogenase polypeptide.

1 58. The organism according to claim 57, said organism dehydrogenating a
2 hydroxyalkyl group of a substrate having the structure:



4 wherein

5 R¹¹, R¹², R¹³ and R¹⁴ are independently selected from H and OH and at least
6 one of R¹¹, R¹², R¹³ and R¹⁴ is OH;

7 n and s are independently selected from the numbers 0 to 16; and

8 p and t are independently selected from 0 to 6, wherein at least one of p and t
9 must be at least one and p + t ≤ 6,

10 said hydroxyalkyl group being dehydrogenated to a member selected from a
11 carboxylic acid, a ketone carbonyl and an aldehyde carbonyl.

1 59. The organism according to claim 50, further comprising an improved
2 solvent resistance polypeptide that confers upon said organism a resistance to an organic

3 solvent that is improved relative to that conferred by a naturally occurring solvent resistance-
4 conferring polypeptide.

1 60. The organism according to claim 59, wherein said improved solvent
2 resistance polypeptide imparts to the organism a resistance to one or more organic
3 compounds selected from olefins, α -hydroxycarboxylic acids, diols, aldehydes, ketones,
4 halogenated hydrocarbons, perfluorocarbons, esters, aryl compounds, carboxylic acids,
5 alcohols, ethers and combinations thereof.

1 61. The organism of claim 59, wherein said improved solvent resistance
2 polypeptide imparts to the organism a resistance to said solvent, wherein the solvent is
3 present in a medium at hypersaturating concentrations.

1 62. The organism according to claim 50, wherein said organism further
2 comprises an epoxide hydrolase polypeptide that exhibits one or more properties improved
3 relative to a corresponding property of a naturally occurring epoxide hydrolase polypeptide.

1 63. The organism according to claim 50, wherein said microorganism
2 further comprises an epoxide isomerase polypeptide that exhibits one or more properties
3 improved relative to a corresponding property of a naturally occurring epoxide isomerase
4 polypeptide.

1 64. The organism of claim 50, wherein said organism further comprises two
2 or more recombinant polynucleotides selected from the group consisting of
3 an improved transferase polypeptide that exhibits one or more
4 properties improved relative to a corresponding property of a naturally occurring transferase
5 polypeptide;
6 an improved epoxide hydrolase peptide that exhibits one or more
7 properties improved relative to a corresponding property of a naturally occurring epoxide
8 hydrolase polypeptide;
9 an improved ligase peptide that exhibits one or more properties
10 improved relative to a corresponding property of a naturally occurring ligase polypeptide;
11 an improved racemase polypeptide that exhibits one or more properties
12 improved relative to a corresponding property of a naturally occurring racemase polypeptide;

13 an improved dehydrogenase polypeptide that exhibits one or more
14 properties improved relative to a corresponding property of a naturally occurring
15 dehydrogenase polypeptide;
16 an improved epoxide isomerase polypeptide that exhibits one or more
17 properties improved relative to a corresponding property of a naturally occurring epoxide
18 isomerase polypeptide; and
19 an improved solvent resistance polypeptide that confers upon said
20 organism a resistance to an organic solvent that is improved relative to that conferred by a
21 naturally occurring solvent resistance-conferring polypeptide.

1 65. A method for preparing an epoxide group, said method comprising
2 contacting a substrate comprising a carbon-carbon double bond with an organism according
3 to claim 50, thereby forming said epoxide group.

1 66. The method according to claim 65, wherein said substrate is selected
2 from styrene, styrene substituted on the phenyl group, divinylbenzene, divinylbenzene
3 substituted on the phenyl group, isoprene, butadiene, diallyl ether, allyl phenyl ether, allyl
4 phenyl ether substituted on the phenyl group, allyl alkyl ether, allyl aralkyl ether,
5 vinylcyclohexene, vinylnorbornene, and acrolein.

1 67. A method for converting an olefin into a vicinal diol, said method
2 comprising:
3 (a) contacting said olefin with an organism according to claim 50 to form an
4 epoxide; and
5 (b) contacting said epoxide with an organism comprising an epoxide
6 hydrolase polypeptide, thereby forming said vicinal diol.

1 68. The method according to claim 67, wherein said epoxide hydrolase
2 polypeptide exhibits one or more properties improved relative to corresponding properties of
3 a naturally occurring epoxide hydrolase polypeptide.

1 69. The method according to claim 67, wherein said polypeptide of (a)
2 and said polypeptide of (b) are expressed in the same host cell.

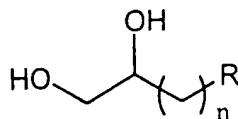
1 70. The method according to claim 67, further comprising,

2 (c) contacting said vicinal diol with an organism comprising a polypeptide
3 selected from a ligase polypeptide and a transferase polypeptide, thereby forming a vicinal
4 diol adduct.

1 71. The method according to claim 70, wherein said polypeptide of (c) is
2 a polypeptide exhibiting one or more properties improved over a corresponding property of
3 an analogous naturally occurring polypeptide.

1 72. The method according to claim 70, wherein said polypeptide of (a),
2 said polypeptide of (b) and said polypeptide of (c) are expressed in the same host cell.

1 73. The method according to claim 67, wherein said vicinal diol has the
2 structure:



4 wherein

5 R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl,
6 heterocyclyl, substituted heterocyclyl, $-NR^2R^3$, $-OR^2$, $-CN$,
7 $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups,

8 R^2 and R^3 are members independently selected from H, alkyl, substituted
9 alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl,
10 heterocyclyl and substituted heterocyclyl groups;

11 R^4 is selected from $=O$ and $=S$, and

12 n is a number between 0 and 10, inclusive.

1 74. The method according to claim 73, wherein

2 R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl
3 $-NR^2R^3$, $-OR^2$, $-CN$, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups,

4 R^2 and R^3 are members independently selected from H, alkyl, substituted
5 alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl,
6 heterocyclyl and substituted heterocyclyl groups; and

7 R^4 is selected from $=O$ and $=S$.

1 75. A method for converting an olefin into an α -hydroxycarboxylic acid,
2 said method comprising:

3 (a) contacting said olefin with an organism according to claim 50 to form an
4 epoxide;

5 (b) contacting said epoxide with an organism comprising an epoxide
6 hydrolase polypeptide to form a vicinal diol; and

7 (c) contacting said vicinal diol with an organism comprising a dehydrogenase
8 polypeptide to form said α -hydroxycarboxylic acid.

1 76. The method according to claim 75, wherein at least one of said
2 hydrolase polypeptide and said dehydrogenase polypeptide exhibits at least one property
3 improved relative to a corresponding property in an analogous naturally occurring
4 polypeptide.

1 77. The method according to claim 78, wherein said polypeptide of (a), of
2 (b) and of (c) are expressed in the same host cell.

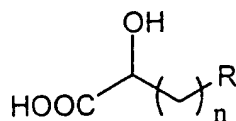
1 78. A method for converting an olefin into an α -hydroxycarboxylic acid,
2 said method comprising, contacting said olefin with an organism according to claim 64,
3 wherein said two or more recombinant polynucleotides are an improved epoxide hydrolase
4 and an improved dehydrogenase.

1 79. The method according to claim 78, further comprising:
2 (d) contacting said α -hydroxycarboxylic acid with an organism comprising an
3 improved polypeptide having an activity selected from ligase, transferase and combinations
4 thereof, thereby forming a α -hydroxycarboxylic acid adduct.

1 80. The method according to claim 79, wherein at least two of said
2 polypeptide of (a), (b), (c), (d) are expressed in the same host cell.

1 81. The method according to claim 79, wherein at least one of said
2 polypeptide selected from ligase, transferase and combinations thereof is an improved
3 polypeptide.

1 82. The method according to claim 78, wherein said α -hydroxycarboxylic
2 acid has the structure:



3
4 wherein

5 R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl,
6 heterocyclyl, substituted heterocyclyl, $-\text{NR}^2\text{R}^3$, $-\text{OR}^2$, $-\text{CN}$,
7 $\text{C}(\text{R}^4)\text{NR}^2\text{R}^3$ and $\text{C}(\text{R}^4)\text{OR}^2$ groups,

8 R^2 and R^3 are members independently selected from H, alkyl, substituted
9 alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl,
10 heterocyclyl and substituted heterocyclyl groups;

11 R^4 is selected from $=\text{O}$ and $=\text{S}$, and

12 n is a number between 0 and 10, inclusive.

1 83. The method according to claim 82 wherein

2 R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl
3 $-\text{NR}^2\text{R}^3$, $-\text{OR}^2$, $-\text{CN}$, $\text{C}(\text{R}^4)\text{NR}^2\text{R}^3$ and $\text{C}(\text{R}^4)\text{OR}^2$ groups,

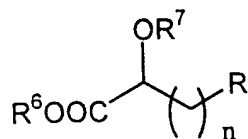
4 R^2 and R^3 are members independently selected from H, alkyl, substituted
5 alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl,
6 heterocyclyl and substituted heterocyclyl groups; and

7 R^4 is selected from $=\text{O}$ and $=\text{S}$.

1 84. The method according to claim 79, wherein said transferase activity is
2 selected from glycosyl transferase activity and methyltransferase activity.

1 85. The method according to claim 84, wherein said methyl transferase is
2 a S-adenosylmethionine dependent O-methyltransferase.

1 86. The method according to claim 79, wherein said α -
2 hydroxycarboxylic acid adduct has the structure:



wherein

R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl, substituted heterocyclyl, $-NR^2R^3(R^4)_m$, $-OR^2$, $-CN$, $C(R^5)NR^2R^3$ and $C(R^5)OR^2$ groups,

R^2 , R^3 and R^4 are members independently selected from said group consisting of H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclyl and substituted heterocyclyl groups;

R^5 is selected from $=O$ and $=S$;

R^6 is selected from H, alkyl and substituted alkyl groups;

R^7 is $C(O)R^8$, wherein R^8 is selected from H alkyl and substituted alkyl groups and R^7 and R^8 are not both H;

m is 0 or 1, such that when m is 1, an ammonium salt is provided; and

n is a number between 0 and 10, inclusive.

87. The method according to claim 86 wherein

R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl $-NR^2R^3$, $-OR^2$, $-CN$, $C(R^5)NR^2R^3$ and $C(R^5)OR^2$ groups

R^2 and R^3 are members independently selected from said group consisting of H, C_1 - C_6 alkyl and allyl; and

R^5 is $=O$.

88. A method for preparing a hydroxy group, said method comprising:

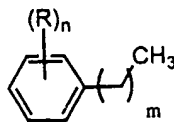
(a) contacting a substrate comprising a terminal methyl group with a microorganism according to claim 50, thereby forming a hydroxymethyl group.

89. The method according to claim 88, wherein said substrate comprises

an alkyl-terminal methyl group as a component of a substrate selected from arylalkyl groups, substituted arylalkyl groups, heteroarylalkyl groups, and substituted heteroarylalkyl groups.

90. The method according to claim 88, wherein said substrate has the

structure



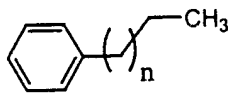
4 wherein,
 5 each of said n R groups is a member selected from the group consisting of H,
 6 alkyl groups and substituted alkyl groups;
 7 m is a number from 0 to 10, inclusive; and
 8 n is a number from 0 to 5, inclusive.

1 91. The method according to claim 90, wherein said substrate comprises
 2 benzene substituted with a member selected from the group of straight-chain alkyl groups
 3 branched-chain alkyl groups and combinations thereof.

1 92. The method according to claim 91, wherein said substrate comprises
 2 benzene substituted with a member selected from C₁-C₆ straight-chain, C₁-C₆ branched-
 3 chain alkyl and combinations thereof.

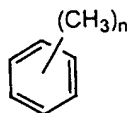
1 93. The method according to claim 92, wherein said alkyl group is
 2 selected from ethyl, *n*-propyl, *i*-propyl, *t*-butyl and combinations thereof.

1 94. The method according to claim 92, wherein said substrate is



2
 3 wherein n is a number between 0 and 9, inclusive.

1 95. The method according to claim 92, wherein said substrate has the
 2 structure:



3
 4 wherein n is a number between 1 and 6, inclusive.

1 96. The method according to claim 88, wherein said hydroxy group is a
 2 component of a member selected from benzyl alcohol, substituted benzyl alcohol, 2-
 3 phenylethanol, substituted 2-phenylethanol, 3-phenylpropanol and substituted 3-
 4 phenylpropanol.

1 **97.** The method according to claim 88, further comprising:
2 (b) contacting said hydroxymethyl group with an organism comprising an
3 acyltransferase, thereby forming an acylated hydroxy adduct.

1 **98.** The method according to claim 97, wherein said acyltransferase
2 exhibits one or more properties improved relative to a corresponding property of a naturally
3 occurring acyltransferase.

1 **99.** The method according to claim 97, wherein said polypeptide of (a)
2 and said polypeptide of (b) are expressed by the same host cell.

1 **100.** The method according to claim 88, further comprising:
2 (b) contacting said hydroxymethyl group with a microorganism comprising
3 an improved glycosyltransferase, thereby forming a glycosylated hydroxy adduct.

1 **101** The method according to claim 100, wherein said glycosyltransferase
2 exhibits one or more properties improved relative to a corresponding property of a naturally
3 occurring glycosyltransferase.

1 **102.** The method according to claim 100, wherein said polypeptide of (a)
2 and said polypeptide of (b) are expressed by the same host cell.

1 **103.** The method according to claim 88, further comprising:
2 (b) contacting said hydroxy group with a microorganism comprising a
3 dehydrogenase, thereby forming a carboxylic acid.

1 **104.** The method according to claim 103, wherein said dehydrogenase
2 exhibits one or more properties improved relative to a corresponding property of a naturally
3 occurring dehydrogenase.

1 **105.** The method according to claim 103, wherein said polypeptide of (a)
2 and said polypeptide of (b) are expressed by the same host cell.

1 **106.** The method according to claim 110, further comprising, contacting
2 said carboxylic acid with a microorganism comprising an improved transferase, thereby
3 forming a carboxylic acid ester.

1 107. A method for preparing a hydroxymethylene group, said method
2 comprising contacting a substrate comprising a methylene group with a microorganism
3 according to claim 50.

1 108. The method according to claim 107, wherein said substrate comprises
2 a member selected from 3,4-dihydrocoumarin and 3,4-dihydrocoumarin residues.

1 109. The method according to claim 107, wherein said substrate is 3,4-
2 dihydrocoumarin and said polypeptide converts said substrate to 4-hydroxy-,4-
3 dihydrocoumarin.

1 110. A method for preparing a hydroxyaryl group, said method comprising:
2 (a) contacting a substrate comprising an aryl group with a microorganism
3 according to claim 50.

1 111. The method according to claim 110, wherein said substrate comprises
2 a group selected from aryl groups, substituted aryl groups, heteroaryl groups and substituted
3 heteroaryl groups.

1 112. The method according to claim 110, further comprising:
2 (b) contacting said hydroxyaryl group with an organism comprising an
3 acyltransferase, thereby forming an acylated hydroxyaryl adduct.

1 113. The method according to claim 112, wherein said acyltransferase
2 exhibits one or more properties improved relative to a corresponding property of a naturally
3 occurring acyltransferase.

1 114. The method according to claim 112, wherein said polypeptide of (a)
2 and said polypeptide of (b) are expressed by the same host cell.

1 115. The method according to claim 112, further comprising:
2 (b) contacting said hydroxyaryl group with a microorganism comprising a
3 glycosyltransferase, thereby forming a glycosylated hydroxyaryl adduct.

1 116 The method according to claim 115, wherein said glycosyltransferase
2 exhibits one or more properties improved relative to a corresponding property of a naturally
3 occurring glycosyltransferase.

1 117. The method according to claim 115, wherein said polypeptide of (a)
2 and said polypeptide of (b) are expressed by the same host cell.

1 118. A screening process comprising:
2 (a) introducing the library of recombinant polynucleotides into a
3 population of test microorganisms such that the recombinant polynucleotides are expressed;
4 (b) placing the organisms in a medium comprising at least one substrate;
5 and
6 (c) and identifying those organisms exhibiting an improved property
7 compared to microorganisms without the recombinant polynucleotide.

1 119. A bioreactor comprising:
2 (a) an improved monooxygenase polypeptide;
3 (b) a redox partner;
4 (c) oxygen;
5 (d) an oxidizable substrate.

1 120. The bioreactor according to claim 119, wherein said polypeptide is
2 immobilized.

1 121. The bioreactor according to claim 119, wherein said polypeptide is a
2 chimeric polypeptide.

1 122. The bioreactor according to claim 119, wherein said polypeptide is a
2 P-450 polypeptide.

1 123. The bioreactor according to claim 122, wherein said P-450 is a
2 peroxide-stable P-450.

1 124. A kit comprising:
2 (a) at least one improved monooxygenase polypeptide; and

3 (b) directions for using said polypeptide to carry out a chemical
4 reaction.

1 125. The kit according to claim 124, wherein said at least one improved
2 monooxygenase polypeptide is a constituent of a library of improved polypeptides.

1 126. A recombinant P450 polypeptide comprising a backbone domain and
2 an active site domain, wherein at least one of said domains comprises at least two contiguous
3 amino acids that are not contiguous in a naturally occurring P450 enzyme.

1 127. The recombinant P450 polypeptide according to claim 126, wherein
2 the junction between the active site domain and the backbone domain is at a location
3 selected from an end of the I helix and within the G-H loop.

1 128. The recombinant P450 polypeptide according to claim 126, wherein
2 the F and G helices are transferred into the backbone P450.

1 129. A polynucleotide that encodes a recombinant P450 polypeptide
2 according to claim 126.

1 130. A method of obtaining a polynucleotide that encodes a recombinant
2 P450 polypeptide comprising a backbone domain and an active site domain, said method
3 comprising:

4 (a) recombining at least first and second forms of a nucleic acid that encodes
5 a P450 active site domain, wherein the first and second forms differ from each other in two
6 or more nucleotides to produce a library of recombinant active site domain encoding
7 polynucleotides; and

8 (b) linking the recombinant active site domain-encoding polynucleotide to a
9 backbone-encoding polynucleotide so that the active site-encoding domain and the
10 backbone-encoding domain are in-frame.

1 131. The method according to claim 130, wherein said backbone is derived
2 from P450_{BMP}.

1 132. The method according to claim 130, wherein said backbone domain
2 and said recombinant active-site domain are joined at a member selected from an end of the I
3 helix and within the G-H loop.

1 133. The method according to claim 130, wherein the F and G helices are
2 transferred into the backbone P450.

1 134. A method of obtaining a polynucleotide that encodes a recombinant
2 P450 polypeptide comprising a backbone domain and an active site domain, said method
3 comprising:

4 (a) recombining at least first and second forms of a nucleic acid that encodes
5 a P450 backbone domain, wherein the first and second forms differ from each other in two
6 or more nucleotides to produce a library of recombinant backbone domain encoding
7 polynucleotides; and

8 (b) linking the recombinant backbone domain-encoding polynucleotide to a
9 active site-encoding polynucleotide so that the backbone-encoding domain and the active
10 site-encoding domain are in-frame.

1 135. The method according to claim 134, wherein said backbone is derived
2 from P450_{BMP}.

1 136. The method according to claim 134, wherein said backbone domain
2 and said recombinant active-site domain are joined at a member selected from an end of the I
3 helix and within the G-H loop.

1 137. The method according to claim 134, wherein the F and G helices are
2 transferred into the backbone P450.

1 138. A method of obtaining a polynucleotide that encodes a recombinant
2 P450 polypeptide comprising a backbone domain and an active site domain, said method
3 comprising:

4 (a) recombining at least first and second forms of a nucleic acid that encodes
5 a P450 active site domain, wherein the first and second forms differ from each other in two

6 or more nucleotides to produce a library of recombinant active site domain encoding
7 polynucleotides;
8 (b) recombining at least first and second forms of a nucleic acid that encodes
9 a P450 backbone domain, wherein the first and second forms differ from each other in two
10 or more nucleotides to produce a library of recombinant backbone domain encoding
11 polynucleotides; and
12 (c) linking the recombinant active site domain-encoding polynucleotide to the
13 recombinant backbone-encoding polynucleotide so that the recombinant active site-encoding
14 domain and the recombinant backbone-encoding domain are in-frame.

1 139. The method according to claim 138, wherein said backbone is derived
2 from P450_{BMP}.

1 140. The method according to claim 138, wherein said backbone domain
2 and said recombinant active-site domain are joined at a member selected from an end of the I
3 helix and within the G-H loop.

1 141. The method according to claim 138, wherein the F and G helices are
2 transferred into the backbone P450.

I/II

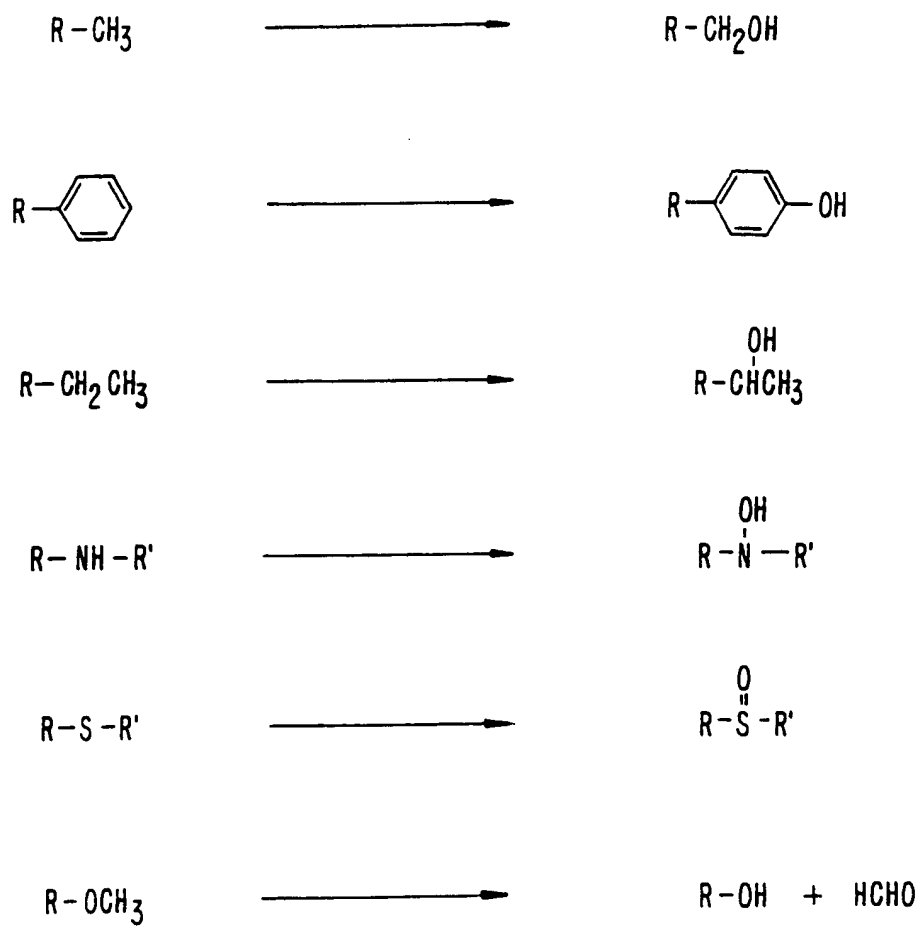


FIG. 1.

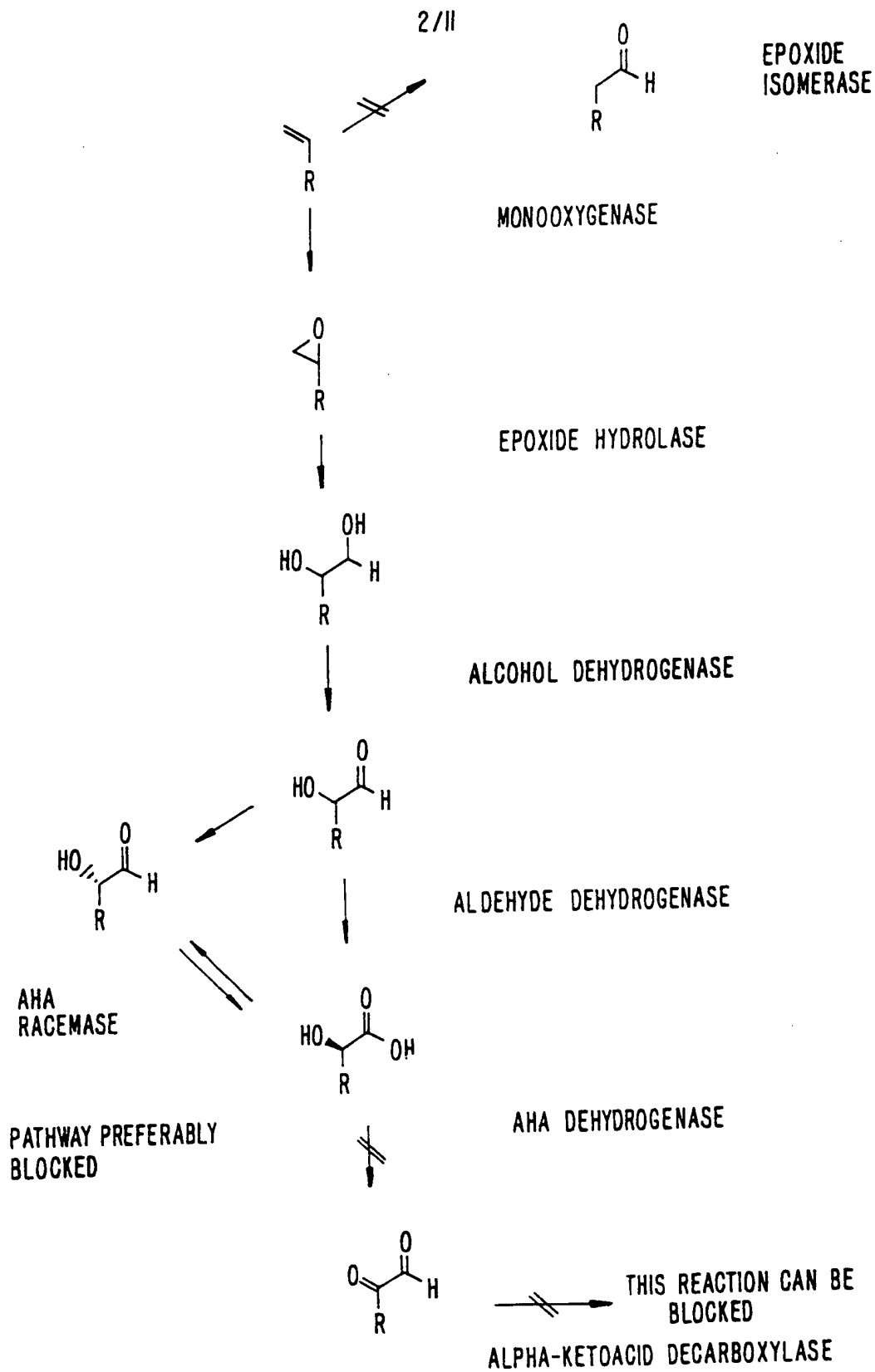


FIG. 2.

SUBSTITUTE SHEET (RULE 26)

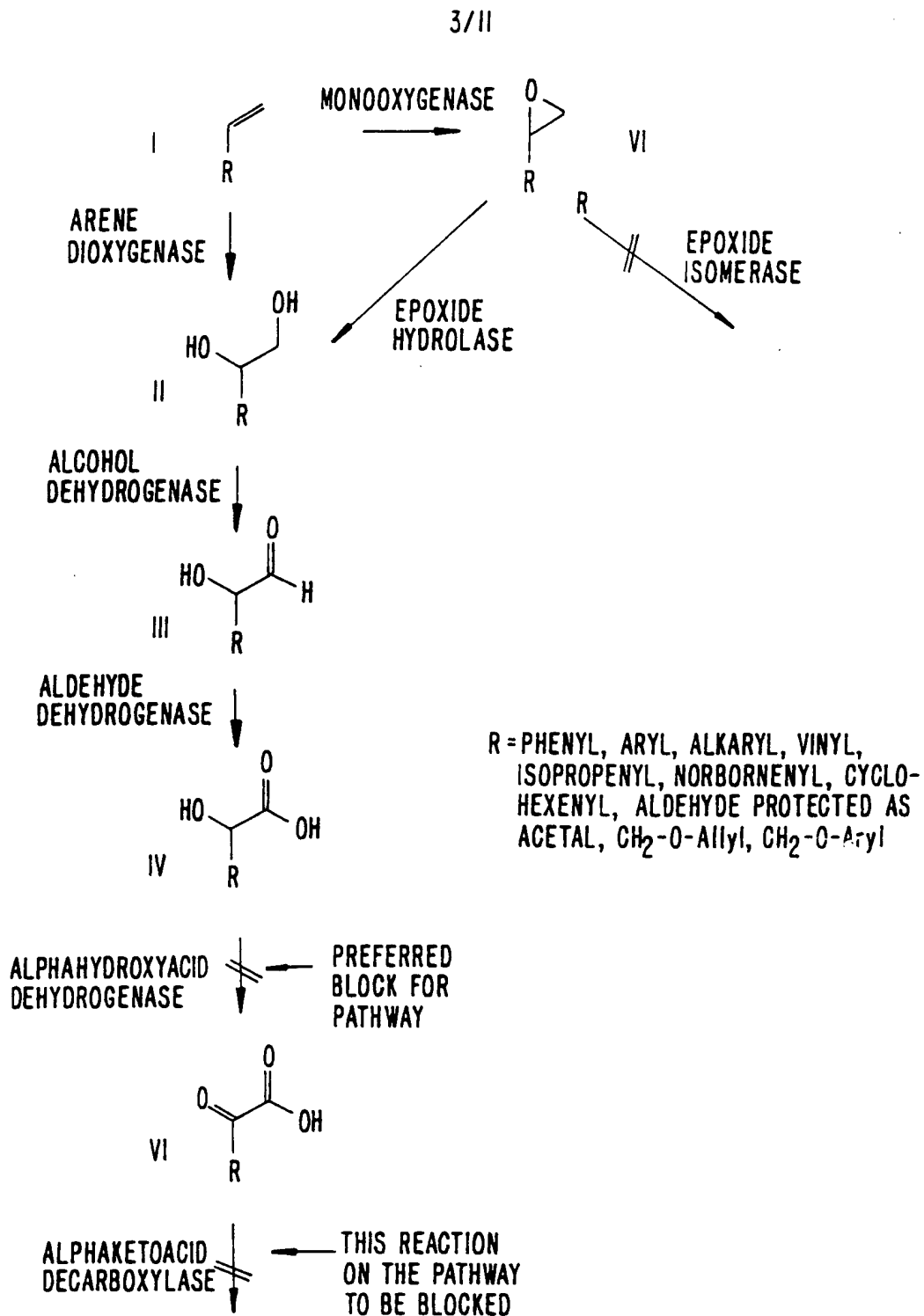


FIG. 3.

4/11

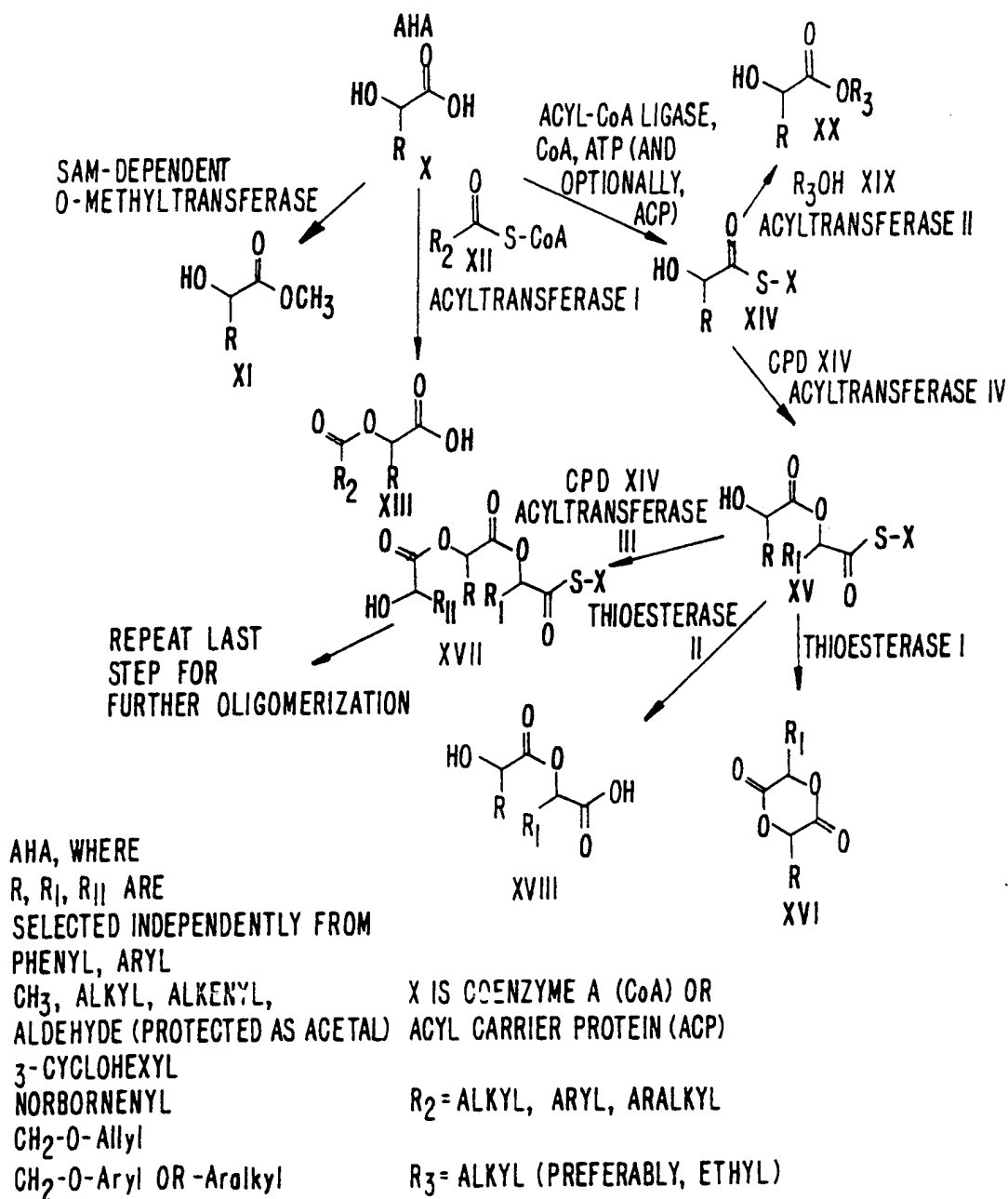
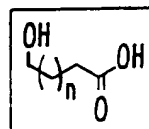
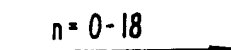
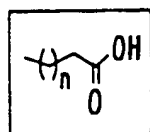


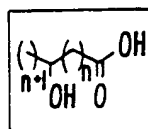
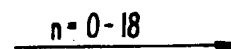
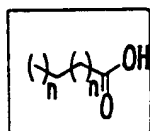
FIG. 4.

5/11

A. FATTY ACIDS

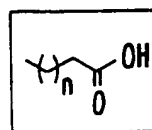
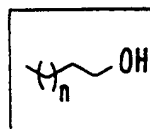
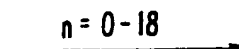
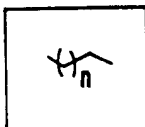


→ DIACIDS



B. n-ALKANES

Fisher-Tropesh Waxe.
n-alkanes
C20-C60



↓

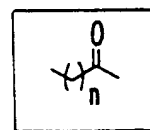
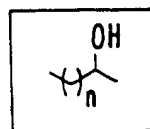
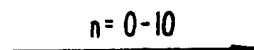
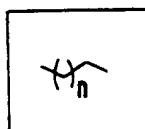


FIG. 5A.

6/11

C. BRANCHED ALKENES AND CARBON BACKBONES

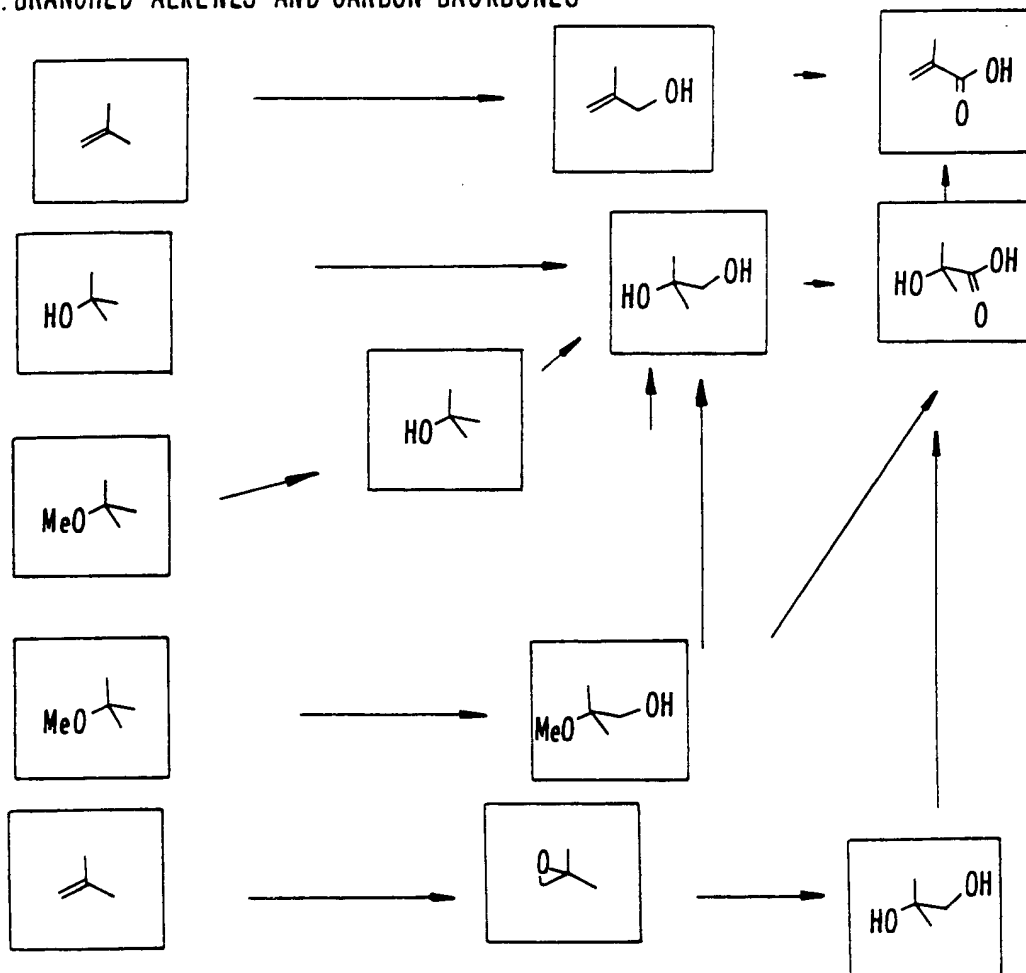


FIG. 5B.

7/11

C. ALICYCLIC COMPOUNDS

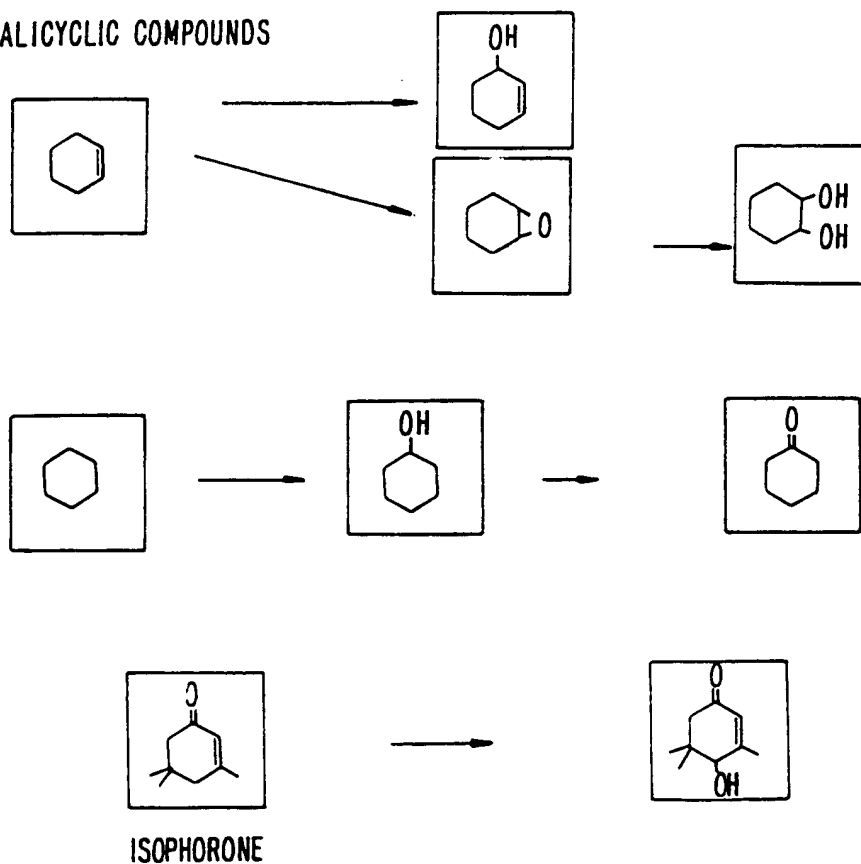


FIG. 5C.

8/11

C. AROMATICS

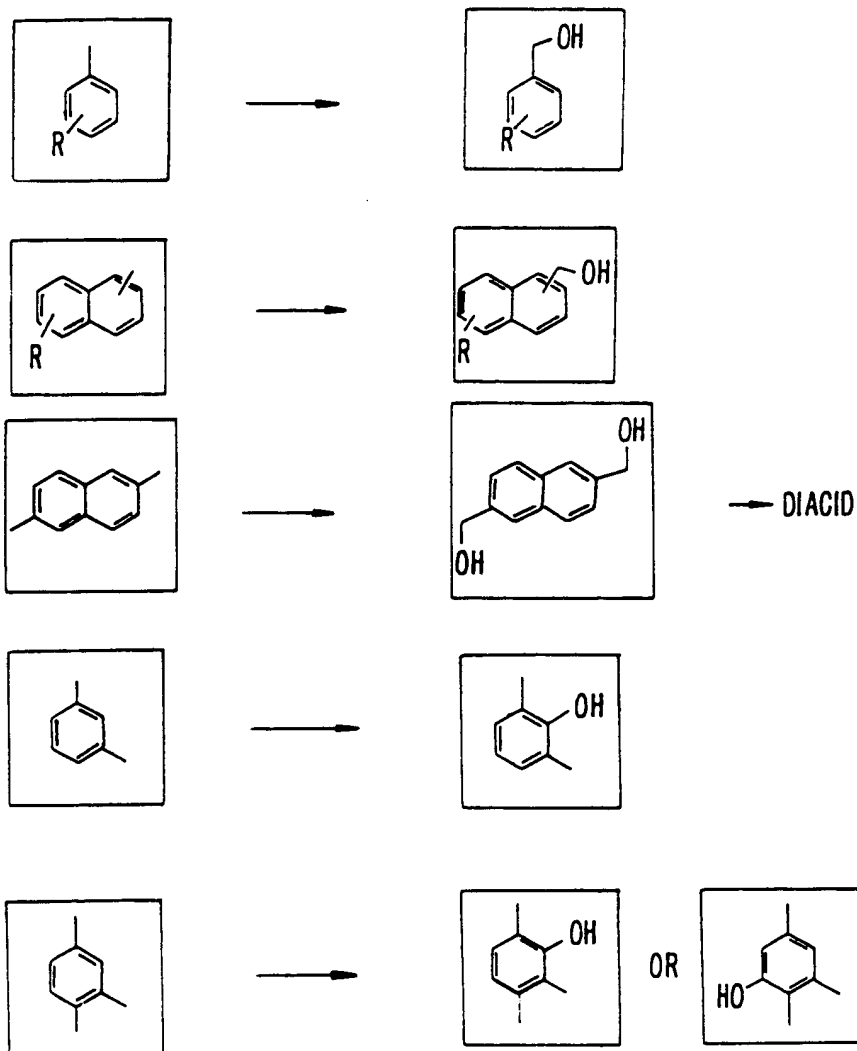


FIG. 5D.

9/11

C. AROMATICS II

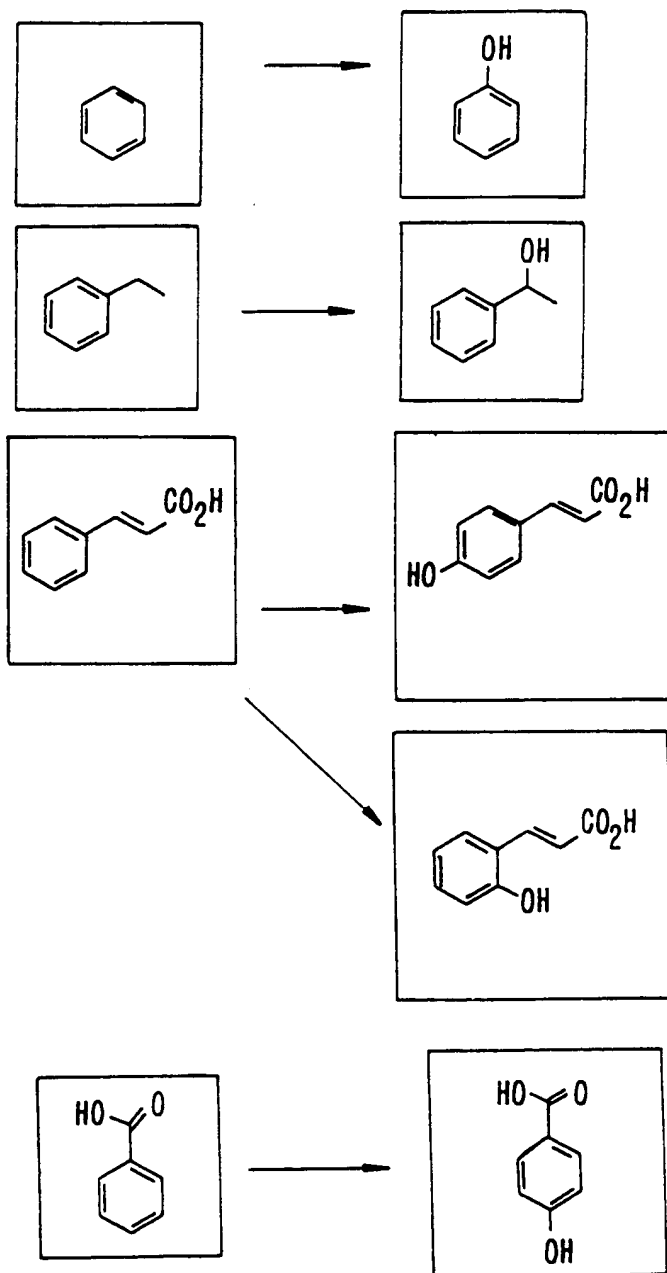


FIG. 5E.

10/11

AROMATICS III

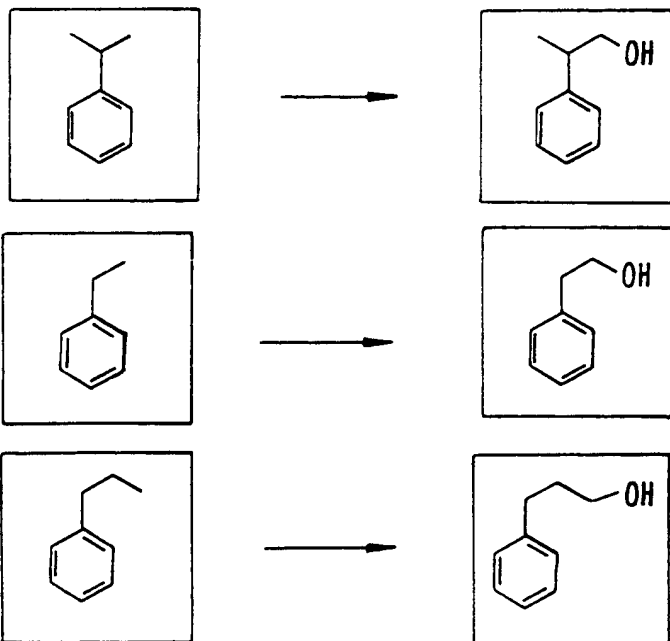
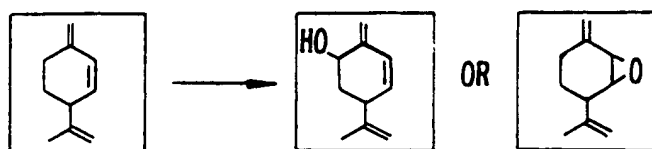
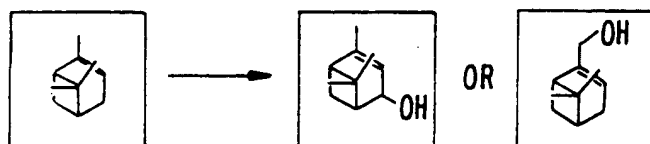
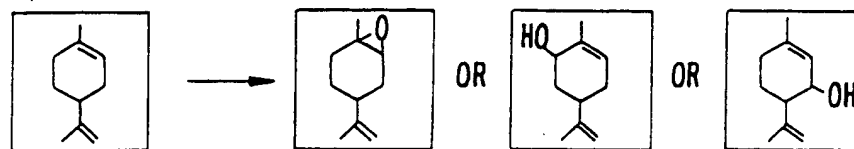


FIG. 5F.

II/II

TERPENOIDS



LINEAR OLEFINS

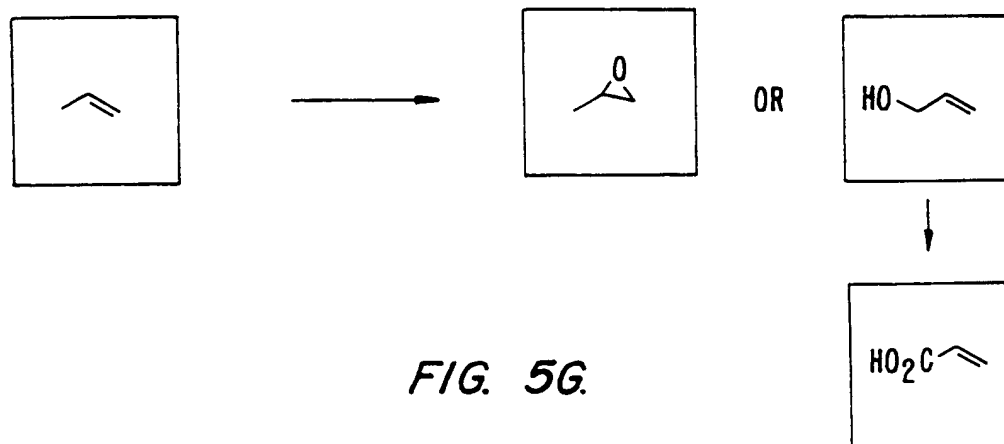


FIG. 5G.

INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 99/18424

A. CLASSIFICATION OF SUBJECT MATTER

IPC 7 C12N15/10 C12N9/02 C12N15/52

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	<p>POMPON D AND NICOLAS A: "Protein engineering by cDNA recombination in yeasts: shuffling of mammalian cytochrome-P450 functions - gene cloning and expression in Saccharomyces cerevisiae; vector construction; mosaic enzyme construction by recombination" GENE, vol. 83, no. 1, 1989, pages 15-24, XP000054542 the whole document</p> <p>---</p> <p>-/--</p>	<p>1-5,7,8, 13-36, 38,50, 118, 124-141</p>

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

2 December 1999

Date of mailing of the international search report

20/12/1999

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Mateo Rosell, A.M.

INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 99/18424

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	<p>WO 97 20078 A (AFFYMAX TECH NV ; CRAMER I ANDREAS (US); STEMMER WILLEM P C (US)) 5 June 1997 (1997-06-05) cited in the application</p> <p>page 6, line 20 -page 8, line 5 page 9, line 10 -page 12, line 29 page 61, line 13-25 page 63, line 26 -page 64, line 8 page 101, line 15 -page 102, line 28; examples 13,15 page 17, line 2-35</p> <p>---</p>	<p>1-5,7,8, 13-36, 38-42, 44, 46-52, 64,79, 81,84, 85,118, 124-141</p>
Y	<p>PANKE S., : "Towards a biocatalyst for (S)-styrene oxide production: characterization of the styrene degradation pathway of Pseudomonas sp. strain VLB120" APPL. ENVIRON. MICROBIOL., vol. 64, no. 6, June 1998 (1998-06) - 2032, page 2043 XP002124400 the whole document</p> <p>---</p>	<p>1-4,7,8, 39-42, 44,46-50</p>
Y	<p>WATANABE I AND SERIZAWA N: "Molecular approaches for production of pravastatin, a HMG-CoA reductase inhibitor: Transcriptional regulation of the cytochrome P450sca gene from Streptomyces carbophilus by ML-236B sodium salt and phenobarbital. " GENE, vol. 210 (1), March 1998 (1998-03), page 109-116 XP004117457 the whole document</p> <p>---</p>	<p>1,2,7, 126-141</p>
Y	<p>WO 98 13485 A (ES HELMUTH H G VAN ; MAXYGEN INC (US); STEMMER WILLEM P C (US)) 2 April 1998 (1998-04-02) cited in the application</p> <p>page 3, line 6-25 page 12, line 10 -page 14, line 21; figure 1</p> <p>---</p>	<p>1,2,13, 14,51, 52,64, 79,81, 84,85</p>
Y	<p>O'KEEFE D P ET AL: "OCCURRENCE AND BIOLOGICAL FUNCTION OF CYTOCHROME P450 MONOOXYGENASES IN THE ACTINOMYCETES" MOLECULAR MICROBIOLOGY,GB,OXFORD, vol. 5, no. 9, 1991, page 2099-2105 XP002913526 the whole document</p> <p>---</p>	<p>1,2,7, 126-141</p>

INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 99/18424

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 95 34679 A (IDLE JEFFREY R ;US HEALTH (US); GONZALEZ FRANK J (US)) 21 December 1995 (1995-12-21) page 1, line 21-30 page 3, line 1 -page 5, line 15 ---	1,2,7, 126-141
Y	AOYAMA T ET AL., : "Cytochrome P-450 H-PCN3 a novel cytochrome P-450 IIIA gene product that is differentially expressed in adult human liver complementary DNA and deduced amino acid sequence and distinct specificities of complementary DNA-expressed HPN1 and H-PCN3 for the metabolism of steroid hormones and cyclosporine" JOURNAL OF BIOLOGICAL CHEMISTRY, vol. 264, no. 18, 1989, pages 10388-10395, XP002124401 the whole document ---	1,2,7, 126-141
Y	DATABASE EMBL NUCLEOTIDE AND PROTEIN SEQUENCES, 10 June 1997 (1997-06-10), XP002124402 HINXTON, GB cited in the application AC= AB004059. Pseudomonas putida dehydrogenase; dioxygenase; ferredoxin; hydratase-aldolase; Iron-sulfur protein large subunit; Iron-sulfur protein small subunit; reductase. abstract ---	1,2,7, 126-141
A	CRAMERI A ET AL: "DNA SHUFFLING OF A FAMILY OF GENES FROM DIVERSE SPECIES ACCELERATES DIRECTED EVOLUTION" NATURE, vol. 391, 1998, page 288-291 XP000775869 ISSN: 0028-0836 cited in the application the whole document ---	1,2
A	RUETTINGER R T ET AL., : "CODING NUCLEOTIDE 5' REGULATORY AND DEDUCED AMINO ACID SEQUENCES OF P-450B-M-3 A SINGLE PEPTIDE CYTOCHROME P-450 NADPH-P-450 REDUCTASE FROM BACILLUS-MEGATERIUM" J BIOL CHEM, vol. 264 (19), 1989, page 10987-10995 XP002124403 the whole document --- -/--	1,2,7, 126-141

INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 99/18424

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>GOTOH O: "SUBSTRATE RECOGNITION SITES IN CYTOCHROME P450 FAMILY 2 CYP2 PROTEINS INFERRED FROM COMPARATIVE ANALYSES OF AMINO ACID AND CODING NUCLEOTIDE SEQUENCES" J BIOL CHEM , vol. 267 (1), 1992, page 83-90 XP002124404 cited in the application the whole document</p>	1,2,7, 126-141
A	<p>LEWIS D F V AND LAKE B G: "Molecular modelling of mammalian CYP2B isoforms and their interaction with substrates, inhibitors and redox partners." XENOBIOTICA , vol. 27 (5), 1997, page 443-478 XP002124405 the whole document</p>	1,2,7, 126-141
A	<p>SEVRIOUKOVA I F AND PETERSON J A: "NADPH-P-450 reductase: Structural and functional comparisons of the eukaryotic and prokaryotic isoforms." BIOCHIMIE , vol. 77 (7-8), 1995, page 562-572 XP002124406 abstract page 563 page 570 -page 571</p>	1, 126-141
A	<p>SHIMOJI M ET AL: "DESIGN OF A NOVEL P450: A FUNCTIONAL BACTERIAL-HUMAN CYTOCHROME P450 CHIMERA" BIOCHEMISTRY,US,AMERICAN CHEMICAL SOCIETY. EASTON, PA, vol. 37, no. 25, 1998, page 8848-8852 XP002913528 ISSN: 0006-2960 the whole document</p>	1, 126-141
A	<p>DATABASE WPI Section Ch, Week 199314 Derwent Publications Ltd., London, GB; Class B04, AN 1993-111879 XP002124410 & JP 05 049474 A (AGENCY OF IND SCI & TECHNOLOGY), 2 March 1993 (1993-03-02) abstract</p>	1,119, 122, 126-141

-/--

INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 99/18424

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,X	<p>WACKETT LP: "Directed evolution of new enzymes and pathways for environmental biocatalysis."</p> <p>ANN N Y ACAD SCI, vol. 864, 13 December 1998 (1998-12-13), page 142-52 XP002124407 abstract page 145-150</p> <p>---</p>	1,2,7, 126-141
P,X	<p>DIERKS E.A. ET AL., : "The catalytic site of cytochrome P450A11 (CYP4A11) and its L131F mutant"</p> <p>THE JOURNAL OF BIOLOGICAL CHEMISTRY, vol. 273, no. 36, September 1998 (1998-09), pages 23055-23061, XP002124408 cited in the application the whole document</p> <p>---</p>	1,2,7, 126-141
P,X	<p>VALETTI F. ET AL., : "Engineering multi-domain redoxproteins containing flavodoxin as bio-transformer: preparatory studies by rational design"</p> <p>BIOSENSORS AND BIOELECTRONICS, vol. 13, September 1998 (1998-09), pages 675-685, XP002124409 page 675 -page 676 page 683</p> <p>-----</p>	1,2,7, 126-141



Europäisches Patentamt
European Patent Office
Office européen des brevets



Publication number: **0 537 178 B1**

EUROPEAN PATENT SPECIFICATION

- (12) (45) Date of publication of patent specification: **31.08.94** (51) Int. Cl.⁵: **C12N 15/53, C11B 1/04, A01H 1/04, C12Q 1/68**
- (21) Application number: **91909981.2**
- (22) Date of filing: **16.05.91**
- (66) International application number:
PCT/US91/03288
- (67) International publication number:
WO 91/18985 (12.12.91 91/28)

(54) NUCLEOTIDE SEQUENCE OF SOYBEAN STEAROYL-ACP DESATURASE GENE.

- (30) Priority: **25.05.90 US 529049**
- (43) Date of publication of application:
21.04.93 Bulletin 93/16
- (45) Publication of the grant of the patent:
31.08.94 Bulletin 94/35
- (84) Designated Contracting States:
AT BE CH DE DK ES FR GB GR IT LI LU NL SE
- (56) References cited:
EP-A- 255 377
WO-A-90/12084
NL-A- 8 800 794
- GENE, vol. 72, 1988, Amsterdam NL, pp. 45-50, A. VAN DER KROL**

BIOTECHNOLOGY, vol. 6, August 1988, New York US, pp. 915-922, M.A.W. HINCHEE et al.

- (73) Proprietor: **E.I. DU PONT DE NEMOURS AND COMPANY**
1007 Market Street
Wilmington Delaware 19898 (US)
- (72) Inventor: **HITZ, William, D.**
2501 West 18th Street
Wilmington, DE 19806 (US)
Inventor: **YADAV, Narendra, S.**
127 Jade Drive
Wilmington, DE 19810 (US)
- (74) Representative: **Hildyard, Edward Martin et al**
Frank B. Dehn & Co.
Imperial House
15-19 Kingsway
London WC2B 6UZ (GB)

EP 0 537 178 B1

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid (Art. 99(1) European patent convention).

CHEMICAL ABSTRACTS, vol. 85, no. 17, October 25, 1976, Columbus, Ohio, US; abstract no. 119683, p. 306

JOURNAL OF BIOLOGICAL CHEMISTRY, vol. 257, no. 20, October 25, 1982, BALTIMORE US, T.A. MCKEON & P.K. STUMPF, pp. 12141-12147

TRENDS IN BIOTECHNOLOGY, vol. 5, 1987, Cambridge GB, pp. 40-46, V. KNAUF

BIOTECHNOLOGY, vol. 7, March 1989, New York US, pp. 257-264, S.D. TANKSLEY et al.

PLANT PHYSIOLOGY, vol. 90, 1989, Rockville USA, pp. 760-764, T. CHEESBROUGH

Description

Soybean oil accounts for about 70% of the 14 billion pounds of edible oil consumed in the United States and is a major edible oil worldwide. It is used in baking, frying, salad dressing, margarine, and a multitude of processed foods. In 1987/88 60 million acres of soybean were planted in the U.S. Soybean is the lowest-cost producer of vegetable oil, which is a by-product of soybean meal. Soybean is agronomically well-adapted to many parts of the U.S. Machinery and facilities for harvesting, storing, and crushing are widely available across the U.S. Soybean products are also a major element of foreign trade since 30 million metric tons of soybeans, 25 million metric tons of soybean meal, and 1 billion pounds of soybean oil were exported in 1987/88. Nevertheless, increased foreign competition has led to recent declines in soybean acreage and production. The low cost and ready availability of soybean oil provides an excellent opportunity to upgrade this commodity oil into higher value specialty oils to both add value to soybean crop for the U.S. farmer and enhance U.S. trade.

Soybean oil derived from commercial varieties is composed primarily of 11% palmitic (16:0), 4% stearic (18:0), 24% oleic (18:1), 54% linoleic (18:2) and 7% linolenic (18:3) acids. Palmitic and stearic acids are, respectively, 16- and 18-carbon-long saturated fatty acids. Oleic, linoleic and linolenic are 18-carbon-long unsaturated fatty acids containing one, two and three double bonds, respectively. Oleic acid is also referred to as a monounsaturated fatty acid, while linoleic and linolenic acids are also referred to as polyunsaturated fatty acids. The specific performance and health attributes of edible oils is determined largely by their fatty acid composition.

Soybean oil is high in saturated fatty acids when compared to other sources of vegetable oil and contains a low proportion of oleic acid, relative to the total fatty acid content of the soybean seed. These characteristics do not meet important health needs as defined by the American Heart Association.

More recent research efforts have examined the role that monounsaturated fatty acid plays in reducing the risk of coronary heart disease. In the past, it was believed that monounsaturates, in contrast to saturates and polyunsaturates, had no effect on serum cholesterol and coronary heart disease risk. Several recent human clinical studies suggest that diets high in monounsaturated fat may reduce the "bad" (low-density lipoprotein) cholesterol while maintaining the "good" (high-density lipoprotein) cholesterol. [See Mattson et al. (1985) *Journal of Lipid Research* 26:194-202, Grundy (1986) *New England Journal of Medicine* 314:745-748, and Mensink et al. (1987) *The Lancet* 1:122-125, all collectively herein incorporated by reference.] These results corroborate previous epidemiological studies of people living in Mediterranean countries where a relatively high intake of monounsaturated fat and low consumption of saturated fat correspond with low coronary heart disease mortality. [Keys, A., *Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease*, Cambridge: Harvard University Press, 1980, herein incorporated by reference.] The significance of monounsaturated fat in the diet was further confirmed by international researchers from seven countries at the Second Colloquium on Monounsaturated Fats held February 26, 1987, in Bethesda, MD, and sponsored by the National Heart, Lung and Blood Institutes [Report, *Monounsaturates Use Said to Lower Several Major Risk Factors*, Food Chemical News, March 2, 1987, p. 44, herein incorporated by reference.]

Soybean oil is also relatively high in polyunsaturated fatty acids -- at levels in far excess of our essential dietary requirement. These fatty acids oxidize readily to give off-flavors and result in reduced performance associated with unprocessed soybean oil. The stability and flavor of soybean oil is improved by hydrogenation, which chemically reduces the double bonds. However, the need for this processing reduces the economic attractiveness of soybean oil.

A soybean oil low in total saturates and polyunsaturates and high in monounsaturate would provide significant health benefits to the United States population, as well as, economic benefit to oil processors. Soybean varieties which produce seeds containing the improved oil will also produce valuable meal as animal feed.

Another type of differentiated soybean oil is an edible fat for confectionary uses. More than 2 billion pounds of cocoa butter, the most expensive edible oil, are produced worldwide. The U.S. imports several hundred million dollars worth of cocoa butter annually. The high and volatile prices and uncertain supply of cocoa butter have encouraged the development of cocoa butter substitutes. The fatty acid composition of cocoa butter is 26% palmitic, 34% stearic, 35% oleic and 3% linoleic acids. About 72% of cocoa butter's triglycerides have the structure in which saturated fatty acids occupy positions 1 and 3 and oleic acid occupies position 2. Cocoa butter's unique fatty acid composition and distribution on the triglyceride molecule confer on it properties eminently suitable for confectionary end-uses: it is brittle below 27°C and depending on its crystalline state, melts sharply at 25-30°C or 35-36°C. Consequently, it is hard and non-greasy at ordinary temperatures and melts very sharply in the mouth. It is also extremely resistant to

rancidity. For these reasons, producing soybean oil with increased levels of stearic acid, especially in soybean lines containing higher-than-normal levels of palmitic acid, and reduced levels of unsaturated fatty acids is expected to produce a cocoa butter substitute in soybean. This will add value to oil and food processors as well as reduce the foreign import of certain tropical oils.

5 Only recently have serious efforts been made to improve the quality of soybean oil through plant breeding, especially mutagenesis, and a wide range of fatty acid composition has been discovered in experimental lines of soybean (Table 1). These findings (as well as those with other oilcrops) suggest that the fatty acid composition of soybean oil can be significantly modified without affecting the agronomic performance of a soybean plant. However, there is no soybean mutant line with levels of saturates less than
10 those present in commercial canola, the major competitor to soybean oil as a "healthy" oil.

TABLE 1

Range of Fatty Acid Percentages Produced by Soybean Mutants	
Fatty Acids	Range of Percentages
Palmitic Acid	6-28
Stearic Acid	3-30
Oleic Acid	17-50
Linoleic Acid	35-60
Linolenic Acid	3-12

There are serious limitations to using mutagenesis to alter fatty acid composition. One is unlikely to
25 discover mutations a) that result in a dominant ("gain-of-function") phenotype, b) in genes that are essential for plant growth, and c) in an enzyme that is not rate-limiting and that is encoded by more than one gene. Even when some of the desired mutations are available in soybean mutant lines their introgression into elite lines by traditional breeding techniques will be slow and expensive, since the desired oil compositions in soybean are most likely to involve several recessive genes.

30 Recent molecular and cellular biology techniques offer the potential for overcoming some of the limitations of the mutagenesis approach, including the need for extensive breeding. Particularly useful technologies are: a) seed-specific expression of foreign genes in transgenic plants [see Goldberg et al. (1989) Cell 56:149-160], b) use of antisense RNA to inhibit plant target genes in a dominant and tissue-specific manner [see van der Krol et al. (1988) Gene 72:45-50], c) transfer of foreign genes into elite
35 commercial varieties of commercial oilcrops, such as soybean [Chee et al. (1989) Plant Physiol. 91:1212-1218; Christou et al. (1989) Proc. Natl. Acad. Sci. U.S.A. 86:7500-7504; Hinchey et al. (1988) Bio/Technology 6:915-922; EPO publication 0 301 749 A2], rapeseed [De Block et al. (1989) Plant Physiol. 91:694-701], and sunflower [Everett et al. (1987) Bio/Technology 5:1201-1204], and d) use of genes as restriction fragment length polymorphism (RFLP) markers in a breeding program, which makes introgression of recessive traits into elite lines rapid and less expensive [Tanksley et al. (1989) Bio/Technology
40 7:257-264]. However, application of each of these technologies requires identification and isolation of commercially-important genes.

Oil biosynthesis in plants has been fairly well-studied [see Harwood (1989) in Critical Reviews in Plant Sciences, Vol. 8(1) 1-43]. The biosynthesis of palmitic, stearic and oleic acids occur in the plastids by the
45 interplay of three key enzymes of the "ACP track": palmitoyl-ACP elongase, stearoyl-ACP desaturase and acyl-ACP thioesterase. Stearoyl-ACP desaturase introduces the first double bond on stearoyl-ACP to form oleoyl-ACP. It is pivotal in determining the degree of unsaturation in vegetable oils. Because of its key position in fatty acid biosynthesis it is expected to be an important regulatory step. While the enzyme's natural substrate is stearoyl-ACP, it has been shown that it can, like its counterpart in yeast and mammalian
50 cells, desaturate stearoyl-CoA, albeit poorly [McKeon et al. (1982) J. Biol. Chem. 257:12141-12147]. The fatty acids synthesized in the plastid are exported as acyl-CoA to the cytoplasm. At least three different glycerol acylating enzymes (glycerol-3-P acyltransferase, 1-acylglycerol-3-P acyltransferase and diacylglycerol acyltransferase) incorporate the acyl moieties from the cytoplasm into triglycerides during oil biosynthesis. These acyltransferases show a strong, but not absolute, preference for incorporating saturated
55 fatty acids at positions 1 and 3 and monounsaturated fatty acid at position 2 of the triglyceride. Thus, altering the fatty acid composition of the acyl pool will drive by mass action a corresponding change in the fatty acid composition of the oil. Furthermore, there is experimental evidence that, because of this specificity, given the correct composition of fatty acids, plants can produce cocoa butter substitutes [Bafor

et al. (1990) JAOCS 67:217-225].

Based on the above discussion, one approach to altering the levels of stearic and oleic acids in vegetable oils is by altering their levels in the cytoplasmic acyl-CoA pool used for oil biosynthesis. There are two ways of doing this genetically: a) altering the biosynthesis of stearic and oleic acids in the plastid
5 by modulating the levels of stearoyl-ACP desaturase in seeds through either overexpression or antisense inhibition of its gene, and b) converting stearoyl-CoA to oleoyl-CoA in the cytoplasm through the expression of the stearoyl-ACP desaturase in the cytoplasm.

In order to use antisense inhibition of stearoyl-ACP desaturase in the seed, it is essential to isolate the gene(s) or cDNA(s) encoding the target enzyme(s) in the seed, since antisense inhibition requires a high-
10 degree of complementarity between the antisense RNA and the target gene that is expected to be absent in stearoyl-ACP desaturase genes from other species or even in soybean stearoyl-ACP desaturase genes that are not expressed in the seed.

The purification and nucleotide sequences of mammalian microsomal stearoyl-CoA desaturases have been published [Thiede et al. (1986) J. Biol. Chem. 262:13230-13235; Ntambi et al. (1988) J. Biol. Chem.
15 263:17291-17300; Kaestner et al. (1989) J. Biol. Chem. 264:14755-14761]. However, the plant enzyme differs from them in being soluble, in utilizing a different electron donor, and in its substrate-specificities. The purification and the nucleotide sequences for animal enzymes do not teach how to purify the plant enzyme or isolate a plant gene. The purification of stearoyl-ACP desaturase was reported from safflower seeds [McKeon et al. (1982) J. Biol. Chem. 257:12141-12147]. However, this purification scheme was not
20 useful for soybean, either because the desaturases are different or because of the presence of other proteins such as the soybean seed storage proteins in seed extracts.

The rat liver stearoyl-CoA desaturase protein has been expressed in *E. coli* [Strittmatter et al. (1988) J. Biol. Chem. 263:2532-2535] but, as mentioned above, its substrate specificity and electron donors are quite distinct from that of the plant.

25

SUMMARY OF THE INVENTION

A means to control the levels of saturated and unsaturated fatty acids in edible plant oils has been discovered. Utilizing the soybean seed stearoyl-ACP desaturase cDNA for either the precursor or enzyme,
30 chimeric genes are created and may be utilized to transform various plants to modify the fatty acid composition of the oil produced. Specifically, one aspect of the present invention is a nucleic acid fragment comprising a nucleotide sequence encoding the soybean seed stearoyl-ACP desaturase cDNA corresponding to the nucleotides 1 to 2243 in SEQ ID NO:1, or any nucleic acid fragment substantially homologous therewith. Preferred are those nucleic acid fragments encoding the soybean seed stearoyl-ACP desaturase
35 precursor or the mature soybean seed stearoyl-ACP desaturase enzyme.

Another aspect of this invention involves a chimeric gene capable of transforming a soybean plant cell comprising a nucleic acid fragment encoding the soybean seed stearoyl-ACP desaturase cDNA operably linked to suitable regulatory sequences producing antisense inhibition of soybean seed stearoyl-ACP desaturase in the seed. Preferred are those chimeric genes which incorporate nucleic acid fragments
40 encoding the soybean seed stearoyl-ACP desaturase precursor or the mature soybean seed stearoyl-ACP desaturase enzyme.

Yet another embodiment of the invention involves a method of producing seed oil containing modified levels of saturated and unsaturated fatty acids comprising: (a) transforming a plant cell with a chimeric gene described above, (b) growing sexually mature plants from said transformed plant cells, (c) screening
45 progeny seeds from said sexually mature plants for the desired levels of stearic acid, and (d) crushing said progeny seed to obtain said oil containing modified levels of stearic acid. Preferred plant cells and oils are derived from soybean, rapeseed, sunflower, cotton, cocoa, peanut, safflower, and corn. Preferred methods of transforming such plant cells would include the use of Ti and Ri plasmids of *Agrobacterium*, electroporation, and high-velocity ballistic bombardment.

50

DETAILED DESCRIPTION OF THE INVENTION

The present invention describes a nucleic acid fragment that encodes soybean seed stearoyl-ACP desaturase. This enzyme catalyzes the introduction of a double bond between carbon atoms 9 and 10 of
55 stearoyl-ACP to form oleoyl-ACP. It can also convert stearoyl-CoA into oleoyl-CoA, albeit with reduced efficiency. Transfer of the nucleic acid fragment of the invention, or a part thereof that encodes a functional enzyme, with suitable regulatory sequences into a living cell will result in the production or over-production of stearoyl-ACP desaturase, which in the presence of an appropriate electron donor, such as ferredoxin,

may result in an increased level of unsaturation in cellular lipids, including oil, in tissues when the enzyme is absent or rate-limiting.

Occasionally, reintroduction of a gene or a part thereof into a plant results in the inhibition of both the reintroduced and the endogenous gene, Jorgenson (December, 1990) Trends in Biotechnology 340-344. Therefore, reintroduction of the nucleic acid fragment of the invention is also expected to, in some cases, result in inhibition of the expression of endogenous seed stearyl-ACP desaturase and would then result in increased level of saturation in seed oil.

Transfer of the nucleic acid fragment of the invention into a soybean plant with suitable regulatory sequences that transcribe the antisense RNA complementary to the mRNA, or its precursor, for seed stearyl-ACP desaturase may result in the inhibition of the expression of the endogenous stearyl-ACP desaturase gene and, consequently, in reduced desaturation in the seed oil.

The nucleic acid fragment of the invention can also be used as a restriction fragment length polymorphism marker in soybean genetic studies and breeding programs.

In the context of this disclosure, a number of terms shall be utilized. As used herein, the term "nucleic acid" refers to a large molecule which can be single stranded or double stranded, composed of monomers (nucleotides) containing a sugar, phosphate and either a purine or pyrimidine. A "nucleic acid fragment" is a fraction of a given nucleic acid molecule. In higher plants, deoxyribonucleic acid (DNA) is the genetic material while ribonucleic acid (RNA) is involved in the transfer of the information in DNA into proteins. A "genome" is the entire body of genetic material contained in each cell of an organism. The term "nucleotide sequence" refers to a polymer of DNA or RNA which can be single- or double-stranded, optionally containing synthetic, non-natural or altered nucleotide bases capable of incorporation into DNA or RNA polymers. As used herein, the term "homologous to" refers to the complementarity between the nucleotide sequence of two nucleic acid molecules or between the amino acid sequences of two protein molecules. Estimates of such homology are provided by either DNA-DNA or DNA-RNA hybridization under conditions of stringency as is well understood by those skilled in the art [as described in Hames and Higgins, Eds. (1985) Nucleic Acid Hybridisation, IRL Press, Oxford, U.K.]; or by the comparison of sequence similarity between two nucleic acids or proteins. As used herein, "substantially homologous" refers to nucleic acid molecules which require less stringent conditions of hybridization than those for homologous sequences, and coding DNA sequence which may involve base changes that do not cause a change in the encoded amino acid, or which involve base changes which may alter an amino acid, but not affect the functional properties of the protein encoded by the DNA sequence.

Thus, the nucleic acid fragments described herein include molecules which comprise possible variations of the nucleotide bases derived from deletion, rearrangement, random or controlled mutagenesis of the nucleic acid fragment, and even occasional nucleotide sequencing errors so long as the DNA sequences are substantially homologous.

"Gene" refers to a nucleic acid fragment that expresses a specific protein, including regulatory sequences preceding (5' non-coding) and following (3' non-coding) the coding region. "Stearyl-ACP desaturase gene" refers to a nucleic acid fragment that expresses a protein with stearyl-ACP desaturase activity. "Native" gene refers to the gene as found in nature with its own regulatory sequences. "Chimeric" gene refers to a gene that comprises heterogeneous regulatory and coding sequences. "Endogenous" gene refers to the native gene normally found in its natural location in the genome. A "foreign" gene refers to a gene not normally found in the host organism but that is introduced by gene transfer.

"Coding sequence" refers to a DNA sequence that codes for a specific protein and excludes the non-coding sequences. It may constitute an "uninterrupted coding sequence", i.e., lacking an intron, such as in a cDNA or it may include one or more introns bounded by appropriate splice junctions. An "intron" is a sequence of RNA which is transcribed in the primary transcript but which is removed through cleavage and re-ligation of the RNA within the cell to create the mature mRNA that can be translated into a protein.

"Translation initiation codon" and "translation termination codon" refer to a unit of three adjacent nucleotides in a coding sequence that specifies initiation and chain termination, respectively, of protein synthesis (mRNA translation). "Open reading frame" refers to the amino acid sequence encoded between translation initiation and termination codons of a coding sequence.

"RNA transcript" refers to the product resulting from RNA polymerase-catalyzed transcription of a DNA sequence. When the RNA transcript is a perfect complementary copy of the DNA sequence, it is referred to as the primary transcript or it may be a RNA sequence derived from posttranscriptional processing of the primary transcript and is referred to as the mature RNA. "Messenger RNA" (mRNA) refers to the RNA that is without introns and that can be translated into protein by the cell. "cDNA" refers to a double-stranded DNA that is complementary to and derived from mRNA. "Sense" RNA refers to an RNA transcript that includes the mRNA. "Antisense RNA" refers to an RNA transcript that is complementary to all or part of a

target primary transcript or mRNA and that blocks the expression of a target gene by interfering with the processing, transport and/or translation of its primary transcript or mRNA. The complementarity of an antisense RNA may be with any part of the specific gene transcript, i.e., at the 5' non-coding sequence, 3' non-coding sequence, introns, or the coding sequence. In addition, as used herein, antisense RNA may contain regions of ribozyme sequences that may increase the efficacy of antisense RNA to block gene expression. "Ribozyme" refers to a catalytic RNA and includes sequence-specific endoribonucleases.

As used herein, "suitable regulatory sequences" refer to nucleotide sequences located upstream (5'), within, and/or downstream (3') to a coding sequence, which control the transcription and/or expression of the coding sequences, potentially in conjunction with the protein biosynthetic apparatus of the cell. In artificial DNA constructs, regulatory sequences can also control the transcription and stability of antisense RNA.

"Promoter" refers to a DNA sequence in a gene, usually upstream (5') to its coding sequence, which controls the expression of the coding sequence by providing the recognition for RNA polymerase and other factors required for proper transcription. In artificial DNA constructs promoters can also be used to transcribe antisense RNA. Promoters may also contain DNA sequences that are involved in the binding of protein factors which control the effectiveness of transcription initiation in response to physiological or developmental conditions. It may also contain enhancer elements. An "enhancer" is a DNA sequence which can stimulate promoter activity. It may be an innate element of the promoter or a heterologous element inserted to enhance the level and/or tissue-specificity of a promoter. "Constitutive promoters" refers to those that direct gene expression in all tissues and at all times. "Tissue-specific" or "development-specific" promoters as referred to herein are those that direct gene expression almost exclusively in specific tissues, such as leaves or seeds, or at specific development stages in a tissue, such as in early or late embryogenesis, respectively. "Inducible promoters" refers to those that direct gene expression in response to an external stimulus, such as light, heat-shock and chemical.

The term "expression", as used herein, is intended to mean the production of a functional end-product. In the case of expression or overexpression of the stearoyl-ACP desaturase genes it involves transcription of the gene and translation of the mRNA into precursor or mature stearoyl-ACP desaturase proteins. In the case of antisense inhibition it refers to the production of antisense RNA transcripts capable of preventing the expression of the target protein. "Overexpression" refers to the production of a gene product in transgenic organisms that exceeds levels of production in normal or non-transformed organisms.

The "3' non-coding sequences" refers to that the DNA sequence portion of a gene that contains a polyadenylation signal and any other regulatory signal capable of affecting mRNA processing or gene expression. The polyadenylation signal is usually characterized by affecting the addition of polyadenylic acid tracts to the 3' end of the mRNA precursor.

"Mature" protein refers to a functional desaturase enzyme without its transit peptide. "Precursor" protein refers to the mature protein with a native or foreign transit peptide. "Transit" peptide refers to the amino terminal extension of a polypeptide, which is translated in conjunction with the polypeptide forming a precursor peptide and which is required for its uptake by plastids of a cell.

"Transformation" herein refers to the transfer of a foreign gene into the genome of a host organism and its genetically stable inheritance. "Restriction fragment length polymorphism" refers to different sized restriction fragment lengths due to altered nucleotide sequences in or around variant forms of genes, and may be abbreviated as "RFLP". "Fertile" refers to plants that are able to propagate sexually.

Purification of Soybean Seed Stearoyl-ACP Desaturase

Stearoyl-ACP desaturase protein was purified to near-homogeneity from the soluble fraction of extracts made from developing soybean seeds following its chromatography on Blue Sepharose, anion-exchange, alkyl-ACP sepharose, and chromatofocussing on Mono P (Pharmacia). Because of the lability of the enzyme during purification, the nearly homogenous preparation is purified only ca. a few hundred-fold; the basis of this lability is not understood. Chromatofocussing resolved the enzyme into two peaks of activity: the peak that eluted earlier, with an apparent pI of ca. 6, had a higher specific-activity than the peak eluting later, with an apparent pI of ca. 5.7. The native molecular weight of the purified enzyme was estimated by gel filtration to be ca. 65 kD. SDS-polyacrylamide gel electrophoresis (SDS-PAGE) of the purified desaturase preparation showed it to be a polypeptide of ca. 38 kD, which suggests that the native enzyme is a dimer. A smaller polypeptide is occasionally observed in varying amounts resulting in a doublet in some preparations. This appears to be due to a proteolytic breakdown of the larger one, since the level of the smaller one increases during storage. However, it cannot be ruled out that the enzyme could also be a heterodimer or that there are different-sized isozymes.

A highly purified desaturase preparation was resolved on SDS-PAGE, electrophoretically transferred onto Immobilon®-P membrane (Millipore), and stained with Coomassie blue. The ca. 38 kD protein on the Immobilon®-P was cut out and used to make polyclonal antibody in mice.

A C₄ reverse-phase HPLC column was used to further purify the enzyme that eluted earlier in chromatofocussing. The major protein peak was homogeneous for the ca. 38 kD polypeptide. It was used for determining the N-terminal sequence: Arg-Ser-Gly-Ser-Lys-Glu-Val-Glu-Asn-Ile-Lys-Lys-Pro-Phe-Thr-Pro (SEQ ID NO:3).

Cloning of Soybean Seed Stearoyl-ACP Desaturase cDNA

Based on the N-terminal sequence of the purified desaturase protein, a set of eight degenerate 35 nucleotide-long oligonucleotides was designed for use as a hybridization probe. The design took into account the codon usage in selected soybean seed genes and used five deoxyinosines at selected positions of ambiguity. The probe, following radiolabeling, was used to screen a cDNA expression library made in Lambda ZAP vector from poly A⁺ RNA from 20-day old developing soybean seeds. Six positively-hybridizing plaques were subjected to plaque purification. Sequences of the pBluescript (Stratagene) vector, including the cDNA inserts, from each of six purified phages were excised in the presence of a helper phage and the resultant phagemids used to infect *E. coli* cells resulting in a double-stranded plasmids, pDS1 to pDS6.

The cDNA insert in plasmid pDS1 is flanked at one end (the 5' end of the coding sequence) by the unique Eco RI site and at its other end by the unique Hind III site. Both Eco RI and the Hind III sites are from the vector, pBluescript. The nucleotide sequence of the cDNA insert in pDS1 revealed an open reading frame for 402 amino acids that included the mature protein's N-terminal sequence 43 amino acid residues from the N-terminus of the open reading frame (SEQ ID NO:1). At least part of this "presequence" is the transit peptide required for precursor import into the chloroplast. Although there are four methionines in this presequence that are in-frame with the mature protein sequence, the most likely N-terminal residue is methionine at position -32 (with the N-terminal Arg of mature protein being referred to as +1) since: a) the N-terminal methionine in the transit peptide sequences for all known chloroplast precursor proteins, with only one exception, is followed by alanine, and b) the methionine at position -5 is too close to the N-terminus of the mature protein to be the initiating codon for the transit peptide (the smallest transit sequence found thus far is 31 amino acids long). Thus, it can be deduced that the desaturase precursor protein consists of a 32-amino acid long transit peptide and a 359-amino acid long mature protein. Based on fusion-protein studies in which the C-terminus of foreign proteins is fused either to the desaturase precursor at position -10 (Ser) or to the mature desaturase protein at position +10 (Ile), the N-terminus of a functional stearoyl-ACP desaturase enzyme can range at least ± 10 amino acids from Arg at position +1 (SEQ ID NO:1).

The restriction maps of all six plasmids, though not identical, showed a common 0.7 kb Bgl II fragment found within the coding region of the precursor for stearoyl-ACP desaturase in pDS1. This strongly suggests that all six clones encode the stearoyl-ACP desaturase. The partial restriction maps of plasmids pDS1, pDS5 and pDS6 appear to be the identical. The inserts in pDS2 and pDS3, which differ in their physical maps from each other as well as from that of pDS1, were partially sequenced. Their partial nucleotide sequences, including 262 nucleotides from the 3' non-coding region, were identical to that in pDS1.

Of the several cDNA clones isolated from the soybean cDNA library using pDS1 as hybridization probe, five were sequenced in the 3' non-coding sequence and their sequences compared to that of SEQ ID NO:1. The results are summarized below:

Clone #	Sequence correspondence to SEQ ID NO:1	Percent Identity
1	1291-1552	100
2	1291-1394	100
3	1285-1552	100
4	1285-1552	100
5	1298-1505	91

Thus, while the claimed sequence (SEQ ID NO:1) most likely represents the predominantly-expressed stearoyl-ACP desaturase gene in soybean seed, there is at least one other stearoyl-ACP desaturase gene that is 91% homologous at the nucleotide level to the claimed sequence. The partial sequence of clone #5

is shown in SEQ ID NO:2.

As expected, comparison of the deduced amino-acid sequences for soybean stearyl-ACP desaturase and the rat microsomal stearyl-CoA desaturases did not reveal any significant homology.

In vitro recombinant DNA techniques were used to make two fusion proteins:

- 5 a) a recombinant plasmid pGEXB that encodes a ca. 66 kD fusion protein consisting of a 28 kD glutathione-S-transferase (GST) protein fused at its C-terminus to the ca. 38 kD desaturase precursor protein at amino acid residue -10 from the N-terminus of the mature enzyme (Arg, +1) (SEQ ID NO:1). Extracts of *E. coli* cells harboring pGEXB, grown under conditions that induce the synthesis of the fusion protein, show stearyl-ACP desaturase activity and expression of a ca. 66 kD fusion protein that cross-
10 reacts with antibody made against soybean stearyl-ACP desaturase and that binds to glutathione-agarose affinity column. The affinity column can be used to purify the fusion protein to near-homogeneity in a single step. The desaturase moiety can be cleaved off in the presence of thrombin and separated from the GST by re-chromatography on the glutathione-agarose column; and
- 15 b) a recombinant plasmid, pNS2, that encodes a ca. 42 kD fusion protein consisting of 4 kD of the N-terminus of β -galactosidase fused at its C-terminus to the amino acid residue at position +10 (Ile) from the N-terminus of the mature desaturase protein (Arg, +1) (SEQ ID NO:1). Extract of *E. coli* cells harboring pNS2 express a ca. 42 kD protein that cross-reacts with antibody made against soybean stearyl-ACP desaturase and show stearyl-ACP desaturase activity.

E. coli (pGEXB) can be used to purify the stearyl-ACP desaturase for use in structure-function studies
20 on the enzyme, in immobilized cells or in extracellular desaturations [see Ratledge et al. (1984) Eds., Biotechnology for the Oils and Fats Industry, American Oil Chemists' Society]. *E. coli* (pNS2) can be used to express the desaturase enzyme in vivo. However, for in vivo function it may be necessary to introduce an electron donor, such as ferredoxin and NADPH:ferredoxin reductase. The ferredoxin gene has been cloned from a higher plant [Smeekens et al. (1985) Nucleic Acids Res. 13:3179-3194] and human ferredoxin has
25 been expressed in *E. coli* [Coghlan et al. (1989) Proc. Natl. Acad. Sci. USA, 86:835-839]. Alternatively, one skilled in the art can express the mature protein in microorganisms using other expression vectors described in the art [Sambrook et al. (1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory Press; Milman (1987) Meth. Enzymol. 153:482-491; Duffaud et al. (1987) Meth. Enzymol. 153:492-507; Weinstock (1987) Meth. Enzymol. 154:156-163; E.P.O. Publication 0 295 959 A2].

30 The fragment of the instant invention may be used, if desired, to isolate substantially homologous stearyl-ACP desaturase cDNAs and genes, including those from plant species other than soybean. Isolation of homologous genes is well-known in the art. Southern blot analysis reveals that the soybean cDNA for the enzyme hybridizes to several, different-sized DNA fragments in the genomic DNA of tomato, rapeseed (*Brassica napus*), soybean, corn (a monocotyledonous plant) and *Arabidopsis* (which has a very
35 simple genome). The Southern blot of corn DNA reveals that the soybean cDNA can also hybridize non-specifically, which may make the isolation of the corn gene more difficult. Although we do not know how many different genes or "pseudogenes" (non-functional genes) are present in any plant, it is expected to be more than one, since stearyl-ACP desaturase is an important enzyme. Moreover, plants that are amphidiploid (that is, derived from two progenitor species), such as soybean, rapeseed (*B. napus*), and
40 tobacco will have genes from both progenitor species.

The nucleic acid fragment of the instant invention encoding soybean seed stearyl-ACP desaturase cDNA, or a coding sequence derived from other cDNAs or genes for the enzyme, with suitable regulatory sequences, can be used to overexpress the enzyme in transgenic soybean as well as other transgenic species. Such a recombinant DNA construct may include either the native stearyl-ACP desaturase gene or
45 a chimeric gene. One skilled in the art can isolate the coding sequences from the fragment of the invention by using and/or creating sites for restriction endonucleases, as described in Sambrook et al. [(1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory Press]. Of particular utility are sites for Nco I (5'-CCATGG-3') and Sph I (5'-GCATGC-3') that allow precise removal of coding sequences starting with the initiating codon ATG. The fragment of invention has a Nco I recognition
50 sequence at nucleotide positions 1601-1606 (SEQ ID NO:1) that is 357 bp after the termination codon for the coding sequence. For isolating the coding sequence of stearyl-ACP desaturase precursor from the fragment of the invention, an Nco I site can be engineered by substituting nucleotide A at position 69 with C. This will allow isolation of the 1533 bp Nco I fragment containing the precursor coding sequence. The expression of the mature enzyme in the cytoplasm is expected to desaturate stearyl-CoA to oleoyl-CoA.
55 For this it may be necessary to also express the mature ferredoxin in the cytoplasm, the gene for which has been cloned from plants [Smeekens et al. (1985) Nucleic Acids Res. 13:3179-3194]. For isolating the coding sequence for the mature protein, a restriction site can be engineered near nucleotide position 164. For example, substituting nucleotide G with nucleotide C at position 149 or position 154 would result in the

creation of Nco I site or Sph I site, respectively. This will allow isolation of a 1453 bp Nco I fragment or a 1448 bp Sph I-Nco I fragment, each containing the mature protein sequence. Based on fusion protein studies, the N-terminus of the mature stearyl-ACP desaturase enzyme is not critical for enzyme activity.

Antisense RNA has been used to inhibit plant target genes in a dominant and tissue-specific manner [see van der Krol et al. (1988) *Gene* 72:45-50; Ecker et al. (1986) *Proc. Natl. Acad. Sci. USA* 83:5372-5376; van der Krol et al. (1988) *Nature* 336:866-869; Smith et al. (1988) *Nature* 334:724-726; Sheehy et al. (1988) *Proc. Natl. Acad. Sci. USA* 85:8805-8809; Rothstein et al. (1987) *Proc. Natl. Acad. Sci. USA* 84:8439-8443; Cornelissen et al. (1988) *Nucl. Acids Res.* 17:833-843; Cornelissen (1989) *Nucl. Acid Res.* 17:7203-7209; Robert et al. (1989) *Plant Mol. Biol.* 13:399-409].

The use of antisense inhibition of the seed enzyme would require isolation of the coding sequence for genes that are expressed in the target tissue of the target plant. Thus, it will be more useful to use the fragment of the invention to screen seed-specific cDNA libraries, rather than genomic libraries or cDNA libraries from other tissues, from the appropriate plant for such sequences. Moreover, since there may be more than one gene encoding seed stearyl-ACP desaturase, it may be useful to isolate the coding sequences from the other genes from the appropriate crop. The genes that are most highly expressed are the best targets for antisense inhibition. The level of transcription of different genes can be studied by known techniques, such as run-off transcription.

For expressing antisense RNA in soybean seed from the fragment of the invention, the entire fragment of the invention (that is, the entire cDNA for soybean stearyl-ACP desaturase from the unique Eco RI to Hind III sites in plasmid pDS1) may be used. There is evidence that the 3' non-coding sequences can play an important role in antisense inhibition [Ch'ng et al. (1989) *Proc. Natl. Acad. Sci. USA* 86:10006-10010]. There have also been examples of using the entire cDNA sequence for antisense inhibition [Sheehy et al. (1988) *Proc. Natl. Acad. Sci. USA* 89:8439-8443]. The Hind III and Eco RI sites can be modified to facilitate insertion of the sequences into suitable regulatory sequences in order to express the antisense RNA.

A preferred host soybean plant for the antisense RNA inhibition of stearyl-ACP desaturase for producing a cocoa butter substitute in soybean seed oil is a soybean plant containing higher-than-normal levels of palmitic acid, such as A19 double mutant, which is being commercialized by Iowa State University Research Foundation, Inc. (315 Beardshear, Ames, Iowa 50011).

A preferred class of heterologous hosts for the expression of the coding sequence of stearyl-ACP desaturase precursor or the antisense RNA are eukaryotic hosts, particularly the cells of higher plants. Particularly preferred among the higher plants are the oilcrops, such as soybean (*Glycine max*), rapeseed (*Brassica napus*, *B. campestris*), sunflower (*Helianthus annuus*), cotton (*Gossypium hirsutum*), corn (*Zea mays*), cocoa (*Theobroma cacao*), and peanut (*Arachis hypogaea*). Expression in plants will use regulatory sequences functional in such plants.

The expression of foreign genes in plants is well-established [De Blaere et al. (1987) *Meth. Enzymol.* 153:277-291]. The origin of promoter chosen to drive the expression of the coding sequence or the antisense RNA is not critical as long as it has sufficient transcriptional activity to accomplish the invention by increasing or decreasing, respectively, the level of translatable mRNA for stearyl-ACP desaturase in the desired host tissue. Preferred promoters include strong plant promoters (such as the constitutive promoters derived from Cauliflower Mosaic Virus that direct the expression of the 19S and 35S viral transcripts [Odell et al. (1985) *Nature* 313:810-812; Hull et al. (1987) *Virology* 86:482-493]), small subunit of ribulose 1,5-bisphosphate carboxylase [Morelli et al. (1985) *Nature* 315:200; Broglie et al. (1984) *Science* 224:838; Herrera-Estrella et al. (1984) *Nature* 310:115; Coruzzi et al. (1984) *EMBO J.* 3:1671; Faciotti et al. (1985) *Bio/Technology* 3:241], maize zein protein [Matzke et al. (1984) *EMBO J.* 3:1525], and chlorophyll a/b binding protein [Lampa et al. (1986) *Nature* 316:750-752].

Depending upon the application, it may be desirable to select inducible promoters and/or tissue- or development-specific promoters. Such examples include the light-inducible promoters of the small subunit of ribulose 1,5-bisphosphate carboxylase genes (if the expression is desired in tissues with photosynthetic function).

Particularly preferred tissue-specific promoters are those that allow seed-specific expression. This may be especially useful, since seeds are the primary source of vegetable oils and also since seed-specific expression will avoid any potential deleterious effect in non-seed tissues. Examples of seed-specific promoters include but are not limited to the promoters of seed storage proteins, which can represent up to 90% of total seed protein in many plants. The seed storage proteins are strictly regulated, being expressed almost exclusively in seeds in a highly tissue-specific and stage-specific manner [Higgins et al. (1984) *Ann. Rev. Plant Physiol.* 35:191-221; Goldberg et al. (1989) *Cell* 56:149-160]. Moreover, different seed storage proteins may be expressed at different stages of seed development.

Expression of seed-specific genes has been studied in great detail [see reviews by Goldberg et al. (1989) Cell 56:149-160 and Higgins et al. (1984) Ann. Rev. Plant Physiol. 35:191-221]. There are currently numerous examples for seed-specific expression of seed storage protein genes in transgenic dicotyledonous plants. These include genes from dicotyledonous plants for bean β -phaseolin [Sengupta-Gopalan et al. (1985) Proc. Natl. Acad. Sci. USA 82:3320-3324; Hoffman et al. (1988) Plant Mol. Biol. 11:717-729], bean lectin [Voelker et al. (1987) EMBO J. 6: 3571-3577], soybean lectin [Okamuro et al. (1986) Proc. Natl. Acad. Sci. USA 83: 8240-8244], soybean kunitz trypsin inhibitor [Perez-Grau et al. (1989) Plant Cell 1:095-1109], soybean β -conglycinin [Beachy et al. (1985) EMBO J. 4:3047-3053; Barker et al. (1988) Proc. Natl. Acad. Sci. USA 85:458-462; Chen et al. (1988) EMBO J. 7:297-302; Chen et al. (1989) Dev. Genet. 10:112-122; Naito et al. (1988) Plant Mol. Biol. 11:109-123], pea vicilin [Higgins et al. (1988) Plant Mol. Biol. 11:683-695], pea convicilin [Newbiggin et al. (1990) Planta 180:461], pea legumin [Shirsat et al. (1989) Mol. Gen. Genetics 215:326]; rapeseed napin [Radke et al. (1988) Theor. Appl. Genet. 75:685-694] as well as genes from monocotyledonous plants such as for maize 15-kD zein [Hoffman et al. (1987) EMBO J. 6:3213-3221], and barley β -hordein [Marris et al. (1988) Plant Mol. Biol. 10:359-366] and wheat glutenin [Colot et al. (1987) EMBO J. 6:3559-3564]. Moreover, promoters of seed-specific genes operably linked to heterologous coding sequences in chimeric gene constructs also maintain their temporal and spatial expression pattern in transgenic plants. Such examples include *Arabidopsis thaliana* 2S seed storage protein gene promoter to express enkephalin peptides in *Arabidopsis* and *B. napus* seeds [Vandekerckhove et al. (1989) Bio/Technology 7:929-932], bean lectin and bean β -phaseolin promoters to express luciferase [Riggs et al. (1989) Plant Sci. 63:47-57], and wheat glutenin promoters to express chloramphenicol acetyl transferase [Colot et al. (1987) EMBO J. 6:3559-3564].

Of particular use in the expression of the nucleic acid fragment of the invention will be the heterologous promoters from several extensively-characterized soybean seed storage protein genes such as those for the Kunitz trypsin inhibitor [Jofuku et al. (1989) Plant Cell 1:1079-1093; Perez-Grain et al. (1989) Plant Cell 1:1095-1109], glycinin [Nielson et al. (1989) Plant Cell 1:313-328], β -conglycinin [Harada et al. (1988) Plant Cell 1:415-425]. Promoters of genes for α - and β -subunits of soybean β -conglycinin storage protein will be particularly useful in expressing the mRNA or the antisense RNA to stearoyl-ACP desaturase in the cotyledons at mid- to late-stages of seed development [Beachy et al. (1985) EMBO J. 4:3047-3053; Barker et al. (1988) Proc. Natl. Acad. Sci. USA 85:458-462; Chen et al. (1988) EMBO J. 7:297-302; Chen et al. (1989) Dev. Genet. 10:112-122; Naito et al. (1988) Plant Mol. Biol. 11:109-123] in transgenic plants, since: a) there is very little position effect on their expression in transgenic seeds, and b) the two promoters show different temporal regulation: the promoter for the α -subunit gene is expressed a few days before that for the β -subunit gene; this is important for transforming rapeseed where oil biosynthesis begins about a week before seed storage protein synthesis [Murphy et al. (1989) J. Plant Physiol. 135:63-69].

Also of particular use will be promoters of genes expressed during early embryogenesis and oil biosynthesis. The native regulatory sequences, including the native promoter, of the stearoyl-ACP desaturase gene expressing the nucleic acid fragment of the invention can be used following its isolation by those skilled in the art. Heterologous promoters from other genes involved in seed oil biosynthesis, such as those for *B. napus* isocitrate lyase and malate synthase [Comai et al. (1989) Plant Cell 1:293-300], *Arabidopsis* ACP [Post-Beittenmiller et al. (1989) Nucl. Acids Res. 17:1777], *B. napus* ACP [Safford et al. (1988) Eur. J. Biochem. 174:287-295], *B. campestris* ACP [Rose et al. (1987) Nucl. Acids Res. 15:7197] may also be used. The partial protein sequences for the relatively-abundant enoyl-ACP reductase and acetyl-CoA carboxylase are published [Slabas et al. (1987) Biochim. Biophys. Acta 877:271-280; Cottingham et al. (1988) Biochim. Biophys. Acta 954: 201-207] and one skilled in the art can use these sequences to isolate the corresponding seed genes with their promoters.

Proper level of expression of stearoyl-ACP mRNA or antisense RNA may require the use of different chimeric genes utilizing different promoters. Such chimeric genes can be transferred into host plants either together in a single expression vector or sequentially using more than one vector.

It is envisioned that the introduction of enhancers or enhancer-like elements into either the native stearoyl-ACP desaturase promoter or into other promoter constructs will also provide increased levels of primary transcription for antisense RNA or in RNA for stearoyl-ACP desaturase to accomplish the inventions. This would include viral enhancers such as that found in the 35S promoter [Odell et al. (1988) Plant Mol. Biol. 10:263-272], enhancers from the opine genes [Fromm et al. (1989) Plant Cell 1:977-984], or enhancers from any other source that result in increased transcription when placed into a promoter operably linked to the nucleic acid fragment of the invention.

Of particular importance is the DNA sequence element isolated from the gene for the α -subunit of β -conglycinin that can confer 40-fold seed-specific enhancement to a constitutive promoter [Chen et al. (1988) EMBO J. 7:297-302; Chen et al. (1989) Dev. Genet. 10:112-122]. One skilled in the art can readily isolate

this element and insert it within the promoter region of any gene in order to obtain seed-specific enhanced expression with the promoter in transgenic plants. Insertion of such an element in any seed-specific gene that is expressed at different times than the β -conglycinin gene will result in expression in transgenic plants for a longer period during seed development.

5 The invention can also be accomplished by a variety of other methods to obtain the desired end. In one form, the invention is based on modifying plants to produce increased levels of stearoyl-ACP desaturase by virtue of having significantly larger numbers of copies of either the wild-type or a stearoyl-ACP desaturase gene from a different soybean tissue in the plants. This may result in sufficient increases in stearoyl-ACP desaturase levels to accomplish the invention.

10 Any 3' non-coding region capable of providing a polyadenylation signal and other regulatory sequences that may be required for the proper expression of the stearoyl-ACP desaturase coding region can be used to accomplish the invention. This would include the native 3' end of the substantially homologous soybean stearoyl-ACP desaturase gene(s), the 3' end from a heterologous stearoyl-ACP desaturase gene, the 3' end from viral genes such as the 3' end of the 35S or the 19S cauliflower mosaic virus transcripts, the 3' end from the opine synthesis genes, the 3' ends of ribulose 1,5-bisphosphate carboxylase or chlorophyll a/b binding protein, or 3' end sequences from any source such that the sequence employed provides the necessary regulatory information within its nucleic acid sequence to result in the proper expression of the promoter/stearoyl-ACP desaturase coding region combination to which it is operably linked. There are numerous examples in the art that teach the usefulness of different 3' non-coding regions.

20 Various methods of transforming cells of higher plants according to the present invention are available to those skilled in the art (see EPO publications 0 295 959 A2 and 0 318 341 A1). Such methods include those based on transformation vectors based on the Ti and Ri plasmids of *Agrobacterium* spp. It is particularly preferred to use the binary type of these vectors. Ti-derived vectors transform a wide variety of higher plants, including monocotyledonous and dicotyledonous plants, such as soybean, cotton and rape [Pacciotti et al. (1985) *Bio/Technology* 3:241; Byrne et al. (1987) *Plant Cell, Tissue and Organ Culture* 8:3; Sukhapinda et al. (1987) *Plant Mol. Biol.* 8:209-216; Lorz et al. (1985) *Mol. Gen. Genet.* 199:178; Potrykus (1985) *Mol. Gen. Genet.* 199:183]. Other transformation methods are available to those skilled in the art, such as direct uptake of foreign DNA constructs [see EPO publication 0 295 959 A2], techniques of electroporation [see Fromm et al. (1986) *Nature (London)* 319:791] or high-velocity ballistic bombardment with metal particles coated with the nucleic acid constructs [see Kline et al. (1987) *Nature (London)* 327:70].

30 Once transformed the cells can be regenerated by those skilled in the art. Of particular relevance are the recently described methods to transform foreign genes into commercially important crops, such as rapeseed [see De Block et al. (1989) *Plant Physiol.* 91:694-701], sunflower [Everett et al. (1987) *Bio/Technology* 5:1201], and soybean [McCabe et al. (1988) *Bio/Technology* 6:923; Hinchee et al. (1988) *Bio/Technology* 6:915; Chee et al. (1989) *Plant Physiol.* 91:1212-1218; Christou et al. (1989) *Proc. Natl. Acad. Sci USA* 86:7500-7504; EPO Publication 0 301 749 A2].

40 The use of restriction fragment length polymorphism (RFLP) markers in plant breeding has been well-documented in the art [see Tanksley et al. (1989) *Bio/Technology* 7:257-264]. The nucleic acid fragment of the invention has been mapped to four different loci on a soybean RFLP map [Tingey et al. (1990) *J. Cell Biochem., Supplement* 14E p. 291, abstract R153]. It can thus be used as a RFLP marker for traits linked to these mapped loci. More preferably these traits will include altered levels of stearic acid. The nucleic acid fragment of the invention can also be used to isolate the stearoyl-ACP desaturase gene from variant (including mutant) soybeans with altered stearic acid levels. Sequencing of these genes will reveal nucleotide differences from the normal gene that cause the variation. Short oligonucleotides designed around these differences may be used as hybridization probes to follow the variation in stearic and oleic acids. Oligonucleotides based on differences that are linked to the variation may be used as molecular markers in breeding these variant oil traits.

55 SEQ ID NO:1 represents the nucleotide sequence of a soybean seed stearoyl-ACP desaturase cDNA and the translation reading frame that includes the open reading frame for the soybean seed stearoyl-ACP desaturase. The nucleotide sequence reads from 5' to 3'. Three letter codes for amino acids are used as defined by the Commissioner, 1114 OG 29 (May 15, 1990) incorporated by reference herein. Nucleotide 1 is the first nucleotide of the cDNA insert after the EcoRI cloning site of the vector and nucleotide 2243 is the last nucleotide of the cDNA insert of plasmid pDS1 which encodes the soybean seed stearoyl-ACP desaturase. Nucleotides 70 to 72 are the putative translation initiation codon, nucleotides 166 to 168 are the codon for the N-terminal amino acid of the purified enzyme, nucleotides 1243 to 1245 are the termination codon, nucleotides 1 to 69 are the 5' untranslated sequence, and nucleotides 1246 to 2243 are the 3' untranslated nucleotides. SEQ ID NO:2 represents the partial sequence of a soybean seed stearoyl-ACP desaturase cDNA. The first and last nucleotides (1 and 216 on clone 5) are read 5' to 3' and represent the

3' non-coding sequence. SEQ ID NO:3 represents the N-terminal sequence of the purified soybean seed stearoyl-ACP desaturase. SEQ ID NO:4 represents the degenerate coding sequence for amino acids 5 through 16 of SEQ ID NO:3. SEQ ID NO:5 represents a complementary mixture of degenerate oligonucleotides to SEQ ID NO:4.

5 The present invention is further defined in the following EXAMPLES, in which all parts and percentages are by weight and degrees are Celsius, unless otherwise stated. It should be understood that these EXAMPLES, while indicating preferred embodiments of the invention, are given by way of illustration only. From the above discussion and these EXAMPLES, one skilled in the art can ascertain the essential characteristics of this invention, and without departing from the scope thereof, can make various changes and modifications of the invention to adapt it to various usages and conditions.

EXAMPLE 1

ISOLATION OF cDNA FOR SOYBEAN SEED STEAROYL-ACP DESATURASE

15 PREPARATION OF [9,10-³H]-STEAROYL-ACP

Purification of Acyl Carrier Protein (ACP) from *E. coli*

20 To frozen *E. coli* cell paste, (0.5 kg of 1/2 log phase growth of *E. coli* B grown on minimal media and obtained from Grain Processing Corp, Muscatine, IA) was added 50 mL of a solution 1 M in Tris, 1 M in glycine, and 0.25 M in EDTA. Ten mL of 1 M MgCl₂ was added and the suspension was thawed in a water bath at 50 °C. As the suspension approached 37 °C it was transferred to a 37 °C bath, made to 10 mM in 2-mercaptoethanol and 20 mg of DNase and 50 mg of lysozyme were added. The suspension was stirred for 25 2 h, then sheared by three 20 second bursts in a Waring Blendor. The volume was adjusted to 1 L and the mixture was centrifuged at 24,000xg for 30 min. The resultant supernatant was centrifuged at 90,000xg for 2 h. The resultant high-speed pellet was saved for extraction of acyl-ACP synthase (see below) and the supernatant was adjusted to pH 6.1 by the addition of acetic acid. The extract was then made to 50% in 2-propanol by the slow addition of cold 2-propanol to the stirred solution at 0 °C. The resulting precipitate was allowed to settle for 2 h and then removed by centrifugation at 16,000xg. The resultant supernatant was 30 adjusted to pH 6.8 with KOH and applied at 2 mL/min to a 4.4 x 12 cm column of DEAE-Sephacel® which had been equilibrated in 10 mM MES, pH 6.8. The column was washed with 10 mM MES, pH 6.8 and eluted with 1 L of a gradient of LiCl from 0 to 1.7 M in the same buffer. Twenty mL fractions were collected and the location of eluted ACP was determined by applying 10 µL of every second fraction to a lane of a 35 native polyacrylamide (20% acrylamide) gel electrophoresis (PAGE). Fractions eluting at about 0.7 M LiCl contained nearly pure ACP and were combined, dialyzed overnight against water and then lyophilized.

Purification of Acyl-ACP Synthase

40 Membrane pellets resulting from the high-speed centrifugation described above were homogenized in 380 mL of 50 mM Tris-Cl, pH 8.0, and 0.5 M in NaCl and then centrifuged at 80,000xg for 90 min. The resultant supernatant was discarded and the pellets resuspended in 50 mM Tris-Cl, pH 8.0, to a protein concentration of 12 mg/mL. The membrane suspension was made to 2% in Triton X-100® and 10 mM in MgCl₂, and stirred at 0 °C for 20 min before centrifugation at 80,000xg for 90 min. The protein in the 45 resultant supernatant was diluted to 5 mg/mL with 2% Triton X-100® in 50 mM Tris-Cl, pH 8.0 and, then, made to 5 mM ATP by the addition of solid ATP (disodium salt) along with an equimolar amount of NaHCO₃. The solution was warmed in a 55 °C bath until the internal temperature reached 53 °C and was then maintained at between 53 °C and 55 °C for 5 min. After 5 min the solution was rapidly cooled on ice and centrifuged at 15,000xg for 15 min. The supernatant from the heat treatment step was loaded directly 50 onto a column of 7 mL Blue Sepharose® 4B which had been equilibrated in 50 mM Tris-Cl, pH 8.0, and 2% Triton X-100. The column was washed with 5 volumes of the loading buffer, then 5 volumes of 0.6 M NaCl in the same buffer and the activity was eluted with 0.5 M KSCN in the same buffer. Active fractions were assayed for the synthesis of acyl-ACP, as described below, combined, and bound to 3 mL settled-volume of hydroxylapatite equilibrated in 50 mM Tris-Cl, pH 8.0, 2% Triton X-100®. The hydroxylapatite was 55 collected by centrifugation, washed twice with 20 mL of 50 mM Tris-Cl, pH 8.0, 2% Triton X-100®. The activity was eluted with two 5 mL washes of 0.5 M potassium phosphate, pH 7.5, 2% Triton X-100®. The first wash contained 66% of the activity and it was concentrated with a 30 kD membrane filtration concentrator (Amicon) to 1.5 mL.

Synthesis of [9,10-³H]-Stearoyl-ACP

A solution of stearic acid in methanol (1 mM, 34.8 μ L) was mixed with a solution of [9,10-³H]stearate (Amersham) containing 31.6 μ Ci of ³H and dried in a glass vial. The ACP preparation described above (1.15 mL, 32 nmoles) was added along with 0.1 mL of 0.1 M ATP, 0.05 mL of 80 mM DTT, 0.1 mL of 8 M LiCl, and 0.2 mL of 13% Triton X-100® in 0.5 M Tris-Cl, pH 8.0, with 0.1 M MgCl₂. The reaction was mixed thoroughly and 0.3 mL of the acyl-ACP synthase preparation was added. After 1 h at 37 °C, a 10 μ L aliquot was taken and dried on a small filter paper disc. The disc was washed extensively with chloroform:methanol:acetic acid (8:2:1, v:v:v) and radioactivity retained on the disc was taken as a measure of stearoyl-ACP. At 1 h about 67% of the ACP had been consumed and the reaction did not proceed further in the next 2 h. The reaction mix was diluted 1 to 4 with 20 mM Tris-Cl, pH 8.0, and applied to a 1 mL DEAE-Sephacel® column equilibrated in the same buffer. The column was washed in sequence with 5 mL of 20 mM Tris-Cl, pH 8.0, 5 mL of 80% 2-propanol in 20 mM Tris-Cl, pH 8.0, and eluted with 0.5 M LiCl in 20 mM Tris-Cl, pH 8.0. The column eluate was passed directly onto a 3 mL column of octyl-sepharose® CL-4B which was washed with 10 mL of 20 mM potassium phosphate, pH 6.8, and then eluted with 35% 2-propanol in 2 mM potassium phosphate, pH 6.8. The eluted volume (5.8 mL) contained 14.27 μ Ci of ³H (49% yield based on ACP). The eluted product was lyophilized and redissolved at a concentration of 24 μ M [³H]stearoyl-ACP at 0.9 mCi/ μ mole.

20 PREPARATION OF ALKYL-ACP AFFINITY COLUMNSynthesis of N-hexadecyliodoacetamide

1-Hexadecylamine (3.67 mmole) was dissolved in 14.8 mL of CH₂Cl₂, cooled to 4 °C, and 2.83 mmoles of iodoacetic anhydride in 11.3 mL of CH₂Cl₂ was added dropwise to the stirred solution. The solution was warmed to room temperature and held for 2 h. The reaction mixture was diluted to about 50 mL with CH₂Cl₂ and washed 3 times (25 mL) with saturated sodium bicarbonate solution and then 2 times with water. The volume of the solution was reduced to about 5 mL under vacuum and passed through 25 mL of silica in diethyl ether. The eluate was reduced to an off-white powder under vacuum. This yielded 820 mg (2.03 mmoles) of the N-hexadecyliodoacetamide (71.8% yield). The 300 MHz ¹H NMR spectra of the product was consistent with the expected structure.

Synthesis of N-Hexadecylacetamido-S-ACP

E. coli ACP prepared as above (10 mg in 2 mL of 50 mM Tris-Cl, pH 7.6) was treated at 37 °C with 50 mM DTT for 2 h. The solution was made to 10% TCA, held at 0 °C for 20 min and centrifuged to pellet. The resultant pellet was washed (2 x 2 mL) with 0.1 M citrate, pH 4.2 and redissolved in 3 mL of 50 mM potassium phosphate buffer. The pH of the ACP solution was adjusted to 7.5 with 1 M KOH and 3 mL of N-hexadecyliodoacetamide (3 mM in 2-propanol) was added. A slight precipitate of the N-hexadecyliodoacetamide was redissolved by warming the reaction mix to 45 °C. The mixture was held at 45 °C for 6 h. SDS-PAGE on 20% acrylamide PAGE gel showed approximately 80% conversion to an ACP species of intermediate mobility between the starting, reduced ACP and authentic palmitoyl-ACP. Excess N-hexadecyliodoacetamide was removed from the reaction mix by 4 extractions (3 mL) with CH₂Cl₂ with gentle mixing to avoid precipitation of the protein at the interface.

45 Coupling of N-Hexadecylacetamido-S-ACP to CNBr-activated Sepharose® 4B

Cyanogen bromide-activated Sepharose® 4B (Pharmacia, 2 g) was suspended in 1 mM HCl and extensively washed by filtration and resuspension in 1 mM HCl and finally one wash in 0.1 M NaHCO₃, pH 8.3. The N-hexadecylacetamido-S-ACP prepared above was diluted with an equal volume of 0.2 M NaHCO₃, pH 8.3. The filtered cyanogen bromide-activated Sepharose® 4B (about 5 mL) was added to the N-hexadecylacetamido-S-ACP solution, the mixture was made to a volume of 10 mL with the 0.1 M NaHCO₃, pH 8.3, and mixed by tumbling at room temperature for 6 h. Protein remaining in solution (Bradford assay) indicated approximately 85% binding. The gel suspension was collected by centrifugation, washed once with the 0.1 M NaHCO₃, pH 8.3, and resuspended in 0.1 M ethanolamine adjusted to pH 8.5 with HCl. The suspension was allowed to stand at 4 °C overnight and then washed by centrifugation and resuspension in 12 mL of 0.1 M acetate, pH 4.0, 0.5 M in NaCl and then 0.1 M NaHCO₃, pH 8.3, 0.5 M in NaCl. The alkyl-ACP Sepharose® 4B was packed into a 1 x 5.5 cm column and washed extensively with 20

mM bis-tris propane-Cl (BTP-Cl), pH 7.2, before use.

STEAROYL-ACP DESATURASE ASSAY

5 Stearoyl-ACP desaturase was assayed as described by McKeon et al. [(1982) J. Biol. Chem. 257:12141-12147] except for using [9,10-³H]-stearoyl-ACP. Use of the tritiated substrate allowed assaying the enzyme activity by release of tritium as water, although the assay based on the tritium release underestimates desaturation by a factor of approximately 4 relative to that observed using ¹⁴C-stearoyl-ACP by the method of McKeon et al. [(1982) J. Biol. Chem. 257:12141-12147], apparently because not all tritium
10 is at carbons 9 and 10. Nevertheless, this modification makes the enzyme assay more sensitive, faster and more reliable. The reaction mix consisted of enzyme in 25 μ L of 230 μ g/mL bovine serum albumin (Sigma), 49 μ g/mL catalase (Sigma), 0.75 mM NADPH, 7.25 μ M spinach ferredoxin, and 0.35 μ M spinach ferredoxin:NADPH⁺ oxidoreductase, 50 mM Pipes, pH 6.0, and 1 μ M [9,10-³H]-stearoyl-ACP (0.9 mCi/ μ mole). All reagents, except for the Pipes buffer, labeled substrate and enzyme extract, were
15 preincubated in a volume of 7.25 μ L at pH 8.0 at room temperature for 10 min before adding 12.75 μ L the Pipes buffer and labeled substrate stocks. The desaturase reaction was usually terminated after 5 min by the addition of 400 μ L 10% trichloroacetic acid and 50 μ L of 10 mg/mL bovine serum albumin. After 5 min on ice, the protein precipitate was removed by centrifugation at 13,000xg for 5 min. An aliquot of 425 μ L was removed from the resultant supernatant and extracted twice with 2 mL of hexane. An aliquot of 375 μ L
20 of the aqueous phase following the second hexane extraction was added to 5 mL of ScintiVerse® Bio HP (Fisher) scintillation fluid and used to determine radioactivity released as tritium.

PURIFICATION OF SOYBEAN SEED STEAROYL-ACP DESATURASE

25 Developing soybean seeds, ca. 20-25 days after flowering, were harvested and stored at -80 °C until use. 300 g of the seeds were resuspended in 600 mL of 50 mM BTP-Cl, pH 7.2, and 5 mM dithiothreitol (DTT) in a Waring Blendor. The seeds were allowed to thaw for a few minutes at room temperature to 4 °C and all of the purification steps were carried out at 4 °C unless otherwise noted. The seeds were homogenized in the blendor three times for 30 s each and the homogenate was centrifuged at 14,000xg for
30 20 min. The resultant supernatant was centrifuged at 100,000xg for 1 h. The resultant high-speed supernatant was applied, at a flow-rate of 5 mL/min to a 2.5 x 20 cm Blue Sepharose® column equilibrated in 10 mM BTP-Cl, pH 7.2, 0.5 mM DTT. Following a wash with 2 column volumes of 10 mM BTP-Cl, pH 7.2, 0.5 mM DTT, the bound proteins were eluted in the same buffer containing 1 M NaCl. The eluting protein peak, which was detected by absorbance at 280 nm, was collected and precipitated with 80% ammonium sulfate. Following collection of the precipitate by centrifugation at 10,000xg for 20 min, its
35 resuspension in 10 mM potassium phosphate, pH 7.2, 0.5 mM DTT, overnight dialysis in the same buffer precipitate, and clarification through a 0.45 micron filter, it was applied to a 10 mm x 25 cm Wide-pore™ PEI (NH₂) anion-exchange column (Baker) at 3 mL/min thoroughly equilibrated in buffer A (10 mM potassium phosphate, pH 7.2). After washing the column in buffer A until no protein was eluted, the column
40 was subjected to elution by a gradient from buffer A at 0 min to 0.25 M potassium phosphate (pH 7.2) at 66 min at a flow rate of 3 mL/min. Three mL fractions were collected. The desaturase activity eluted in fractions 17-25 (the activity peak eluted at ca. 50 mM potassium phosphate). The pooled fractions were made to 60 mL with buffer A and applied at 1 mL/min to a 1 x 5.5 cm alkyl-ACP column equilibrated in buffer A containing 0.5 mM DTT. After washing the bound protein with the start buffer until no protein was
45 eluted, the bound protein was eluted by a gradient from buffer A containing 0.5 mM DTT at 0 min to 0.5 M potassium phosphate, pH 7.2, 0.5 mM DTT at 60 min and 1 M potassium phosphate, pH 7.2, 0.5 mM DTT. Four mL fractions were collected. Fractions 15-23, which contained the enzyme with the highest specific activity, were pooled and concentrated to 3 mL by a 30 kD Centricon® concentrator (Millipore) and desalted in a small column of G-25 Sephadex® equilibrated with 25 mM bis-Tris-Cl, pH 6.7. The desalted sample
50 was applied at 1 mL/min to a chromatofocussing Mono P HR 5/20 (Pharmacia) column equilibrated with 25 mM bis-Tris-Cl, pH 6.7, washed with a column volume of the same buffer, and eluted with 1:10 dilution of Polybuffer 74 (Pharmacia) made to pH 5.0 with HCl. Desaturase activity eluted in two peaks: one in fraction 30 corresponding to a pI of ca. 6.0 and the other in fraction 35, corresponding to a pI of ca. 5.7. The protein in the two peaks were essentially composed of ca. 38 kD polypeptide. The first peak had a higher enzyme
55 specific activity and was used for further characterization as well as for further purification on reverse-phase chromatography.

Mono P fractions containing the first peak of enzyme activity were pooled and applied to a C₄ reverse-phase HPLC column (Vydac) equilibrated with buffer A (5% acetonitrile, 0.1% trifluoroacetic acid) and

eluted at 0.1 mL/min with a gradient of 25% buffer B (100% acetonitrile, 0.1% trifluoroacetic acid) and 75% buffer A at 10 min to 50% buffer B and 50% buffer A at 72.5 min. A single major peak eluted at 41.5% buffer B that also ran as a ca. 38 kD protein based on SDS-PAGE. The protein in the peak fraction was used to determine the N-terminal amino acid sequence on a Applied Biosystems 470A Gas Phase Sequencer. The PTH amino acids were analysed on Applied Biosystems 120 PTH Amino Acid Analyzer.

The N-terminal sequence of the ca. 38 kD polypeptide was determined through 16 residues and is shown in SEQ ID NO:3.

CLONING OF SOYBEAN SEED STEAROYL-ACP DESATURASE cDNA

Based on the N-terminal amino acid sequence of the purified soybean seed stearyl-ACP desaturase (SEQ ID NO:3), amino acids 5 through 16, which are represented by the degenerate coding sequence, SEQ ID NO:4, was chosen to design the complementary mixture of degenerate oligonucleotides (SEQ ID NO:5).

The design took into account the codon bias in representative soybean seed genes encoding Bowman-Birk protease inhibitor [Hammond et al. (1984) J. Biol. Chem. 259:9883-9890], glycinin subunit A-2B-1a [Utsumi et al. (1987) Agric. Biol. Chem. 51:3267-3273], lectin (le-1) [Vodkin et al. (1983) Cell 34:1023-1031], and lipoxygenase-1 [Shibata et al. (1987) J. Biol. Chem. 262:10080-10085]. Five deoxyinosines were used at selected positions of ambiguity.

A cDNA library was made as follows: Soybean embryos (ca. 50 mg fresh weight each) were removed from the pods and frozen in liquid nitrogen. The frozen embryos were ground to a fine powder in the presence of liquid nitrogen and then extracted by Polytron homogenization and fractionated to enrich for total RNA by the method of Chirgwin et al. [Biochemistry (1979) 18:5294-5299]. The nucleic acid fraction was enriched for poly A⁺ RNA by passing total RNA through an oligo-dT cellulose column and eluting the poly A⁺ RNA by salt as described by Goodman et al. [(1979) Meth. Enzymol. 68:75-90]. cDNA was synthesized from the purified poly A⁺ RNA using cDNA Synthesis System (Bethesda Research Laboratory) and the manufacturer's instructions. The resultant double-stranded DNA was methylated by DNA methylase (Promega) prior to filling-in its ends with T4 DNA polymerase (Bethesda Research Laboratory) and blunt-end ligating to phosphorylated Eco RI linkers using T4 DNA ligase (Pharmacia). The double-stranded DNA was digested with Eco RI enzyme, separated from excess linkers by passing through a gel filtration column (Sephacrose CL-4B), and ligated to Lambda ZAP vector (Stratagene) as per manufacturer's instructions. Ligated DNA was packaged into phage using Gigapack packaging extract (Stratagene) according to manufacturer's instructions. The resultant cDNA library was amplified as per Stratagene's instructions and stored at -80 °C.

Following the instructions in Lambda ZAP Cloning Kit Manual (Stratagene), the cDNA phage library was used to infect *E. coli* BB4 cells and plated to yield ca. 80,000 plaques per petri plate (150 mm diameter). Duplicate lifts of the plates were made onto nitrocellulose filters (Schleicher & Schuell). Duplicate lifts from five plates were prehybridized in 25 mL of Hybridization buffer consisting of 6X SSC (0.9 M NaCl, 0.09 M sodium citrate, pH 7.0), 5X Denhardt's [0.5 g Ficoll (Type 400, Pharmacia), 0.5 g polyvinylpyrrolidone, 0.5 g bovine serum albumin (Fraction V; Sigma)], 1 mM EDTA, 1% SDS, and 100 µg/mL denatured salmon sperm DNA (Sigma Chemical Co.) at 45 °C for 10 h. Ten pmol of the hybridization probe (see above) were end-labeled in a 52.5 µL reaction mixture containing 50 mM Tris-Cl, pH 7.5, 10 mM MgCl₂, 0.1 mM spermidine-HCl (pH 7.0), 1 mM EDTA (pH 7.0), 5 mM DDT, 200 µCi (66.7 pmoles) of gamma-labeled AT³²P (New England Nuclear) and 25 units of T4 polynucleotide kinase (New England Biolabs). After incubation at 37 °C for 45 min, the reaction was terminated by heating at 68 °C for 10 min. Labeled probe was separated from unincorporated AT³²P by passing the reaction through a Quick-SpinTM (G-25 Sephadex®) column (Boehringer Mannheim Biochemicals). The purified labeled probe (1.2 x 10⁷ dpm/pmole) was added to the prehybridized filters, following their transfer to 10 mL of fresh Hybridization buffer. Following incubation of the filters in the presence of the probe for 16 h in a shaker at 48 °C, the filters were washed in 200 mL of Wash buffer (6X SSC, 0.1% SDS) five times for 5 min each at room temperature, and then once at 48 °C for 5 min. The washed filters were air dried and subjected to autoradiography on Kodak XAR-2 film in the presence of intensifying screens (Lightening Plus, DuPont Cronex®) at -80 °C overnight. Six positively-hybridizing plaques were subjected to plaque purification as described in Sambrook et al. [(1989) Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory Press]. Following the Lambda ZAP Cloning Kit Instruction Manual (Stratagene), sequences of the pBluescript vector, including the cDNA inserts, from each of six purified phages were excised in the presence of a helper phage and the resultant phagemids were used to infect *E. coli* XL-1 Blue cells resulting in double-stranded plasmids, pDS1 to pDS6. The restriction maps of all six plasmids, though not identical, showed a common 0.7 kb Bgl II fragment found in the desaturase gene (see below).

DNA from plasmids pDS1-pDS6 were made by the alkaline lysis miniprep procedure described in Sambrook et al. [(1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory Press]. The alkali-denatured double-stranded DNAs were sequenced using Sequenase® T7 DNA polymerase (US Biochemical Corp.) and the manufacturer's instructions. The sequence of the cDNA insert in plasmid pDS1 is shown in SEQ ID NO:1.

EXAMPLE 2

EXPRESSION OF SOYBEAN SEED STEAROYL-ACP DESATURASE IN E. COLI

Construction of Glutathione-S-Transferase: Stearoyl-ACP Desaturase Fusion Protein

Plasmid pDS1 was linearized with Hind III enzyme, its ends filled-in with Klenow fragment (Bethesda Research Laboratory) in the presence of 50 μ M each of all four deoxynucleotide triphosphates as per manufacturer's instructions, and extracted with phenol:chloroform (1:1). Phosphorylated Eco RI linkers (New England Biolabs) were ligated to the DNA using T4 DNA ligase (New England Biolabs). Following partial digestion with Bgl II enzyme and complete digestion with excess Eco RI enzyme, the DNA was run on an agarose gel and stained with ethidium bromide. The 2.1 kb DNA fragment resulting from a partial Bgl II and Eco RI digestion was cut out of the gel, purified using USBiobclean™ (US Biochemicals), and ligated to Bam HI and Eco RI cleaved vector pGEX2T [Pharmacia; see Smith et al. (1988) Gene 67:31] using T4 DNA ligase (New England Biolabs). The ligated mixture of DNAs were used to transform *E. coli* XL-1 blue cells (Stratagene). Transformants were picked as ampicillin-resistant cells and the plasmid DNA from several transformants analyzed by digestion with Bam HI and Eco RI double restriction digest, as described by Sambrook et al. [(1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory Press]. Plasmid DNA from one transformant, called pGEXB, showed the restriction pattern expected from the correct fusion. The double-stranded plasmid pGEXB was purified and sequenced to confirm the correct fusion by the Sequenase kit (US Biochemical Corp.). The fusion protein consists of a 28 kD glutathione-S-transferase protein fused at its C-terminus to the desaturase precursor protein at Ser at residue -10 from the N-terminus of the mature enzyme (Arg, +1) (SEQ ID NO:1). Thus, it includes ten amino acids from the transit peptide sequence in addition to the mature protein.

Inducible Expression of the Glutathione-S-Transferase-Stearoyl-ACP Desaturase Fusion Protein

Five mL precultures of plasmids pGEXB and pGEX2T, which were grown overnight at 37°C in LB medium [Sambrook et al. (1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory Press] containing 100 μ g/mL ampicillin, were diluted 1:10 in fresh LB medium containing 100 μ g/mL ampicillin and continued to grow on a shaker at 37°C for another 90 min before adding isopropylthio- β -D-galactoside and ferric chloride to final concentrations of 0.3 mM and 50 μ M, respectively. After an additional 3 h on a shaker at 37°C, the cultures were harvested by centrifugation at 4,000xg for 10 min at 4°C. The cells were resuspended in one-tenth of the culture volume of freshly-made and ice-cold Extraction buffer (20 mM sodium phosphate, pH 8.0, 150 mM NaCl, 5 mM EDTA and 0.2 mM phenylmethylsulfonyl fluoride) and re-centrifuged as above. The resultant cells were resuspended in 1/50 vol of the culture in Extraction buffer and sonicated for three ten-second bursts. The sonicated extracts were made to 1% in Triton X-100 and centrifuged at 8,000xg for 1 min in Eppendorf Micro Centrifuge (Brinkmann Instruments) to remove the cellular debris. The supernatant was poured into a fresh tube and used for enzyme assays, SDS-PAGE analysis and purification of the fusion protein.

Five μ L aliquots of the extracts were assayed for stearoyl-ACP desaturase activity in a 1 min reaction, as described in Example I. The activities [net pmol of stearoyl-ACP desaturated per min per mL of extract; the blank (no desaturase enzyme) activity was 15 pmol/min/mL] are shown below:

Reaction mixture	Net pmol/min/mL
<i>E. coli</i> (pGEX2T)	0
<i>E. coli</i> (pGEXB)	399
<i>E. coli</i> (pGEXB) - NADPH	0
<i>E. coli</i> (pGEXB) - ferredoxin	0
<i>E. coli</i> (pGEXB) - ferredoxin-NADPH reductase	3

These results show that the desaturase enzyme activity is present in the extract of *E. coli* cells containing pGEXB but not in that of cells containing the control plasmid pGEX2T. Furthermore, this activity was dependent on an exogenous electron donor.

Proteins in extracts of *E. coli* cells harboring plasmids pGEX2T or pGEXB were resolved by SDS-PAGE, transferred onto Immobilon®-P (Millipore) and cross-reacted with mouse antibody made against purified soybean stearyl-ACP desaturase, as described by Sambrook et al. [(1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory Press]. The resultant Western blot showed that pGEXB encodes for ca. 64 kD GST-stearyl-ACP desaturase fusion polypeptide, although some lower molecular-weight cross-reacting polypeptides can also be observed, which may represent either a degradation or incomplete synthesis of the fusion protein. It is not known whether the GST-desaturase fusion protein is enzymatically active, since the activity observed may be due to the incomplete fusion by the peptides seen here. The fusion polypeptide was not present in extracts of cells harboring the control plasmid (pGEX2T) nor in extracts of cells harboring pGEXB that were not induced by isopropylthio- β -D-galactoside.

15 Purification of the Glutathione-S-Transferase-Stearyl-ACP Desaturase Fusion Protein

The GST-desaturase fusion protein was purified in a one step glutathione-agarose affinity chromatography under non-denaturing conditions, following the procedure of Smith et al. [Gene (1988) 67:31]. For this, the bacterial cell extract was mixed with 1 mL glutathione-agarose (sulfur-linkage, Sigma), equilibrated with 20 mM sodium phosphate, pH 8.0, 150 mM NaCl, for 10 min at room temperature. The beads were collected by centrifugation at 1000xg for 1 min, and washed three times with 1 mL of 20 mM sodium phosphate, pH 8.0, 150 mM NaCl (each time the beads were collected by centrifugation as described above). The fusion protein was eluted with 5 mM reduced glutathione (Sigma) in 50 mM Tris-Cl, pH 8.0. The proteins in the eluted fraction were analyzed by SDS-PAGE and consisted of mostly pure ca. 64 kD GST-desaturase polypeptide, 28 kD GST and a trace of ca. 38 kD desaturase polypeptide. The fusion polypeptide was cleaved in the presence of thrombin, as described by Smith et al. [Gene (1988) 67:31].

Construction of β -Galactosidase-Stearyl-ACP Desaturase Fusion Protein

Plasmid pDS1 DNA was digested with Ssp I and Pvu I enzymes and the digested DNA fragments were resolved by electrophoresis in agarose. The blunt-ended 2.3 kb Ssp I fragment was cut out of the agarose (Pvu I cleaves a contaminating 2.3 kb Ssp I fragment), purified by USBiobclean™ (US Biochemical Corp.), and ligated to vector plasmid pBluescript SK (-) (Stratagene) that had previously been filled-in with Klenow fragment (Bethesda Research Laboratory) following linearization with Not I enzyme. The ligated DNAs were transformed into competent *E. coli* XL-1 blue cells. Plasmid DNA from several ampicillin-resistant transformants were analysed by restriction digestion. One plasmid, called pNS2, showed the expected physical map. This plasmid is expected to encode a ca. 42 kD fusion protein consisting of 4 kD N-terminal of β -galactosidase fused at its C-terminus to isoleucine at residue +10 from the N-terminus of the mature desaturase protein (Arg, +1) (SEQ ID NO:1). Thus, it includes all but the first 10 amino acids of the mature protein. Nucleotide sequencing has not been performed on pNS2 to confirm correct fusion.

Five mL of preculture of *E. coli* cells harboring plasmid pNS2 grown overnight in LB medium containing 100 μ g/mL ampicillin was added to 50 mL of fresh LB medium with 100 μ g/mL ampicillin. After additional 1 h of growth at 37°C in a shaker, isopropylthio- β -D-galactoside and ferric chloride were added to final concentrations of 0.3 mM and 50 μ M, respectively. After another 2 h on a shaker at 37°C, the culture was harvested by centrifugation at 4,000xg for 10 min at 4°C. The cells were resuspended in 1 mL of freshly-made and ice-cold TEP buffer (100 mM Tris-Cl, pH 7.5, 10 mM EDTA and 0.1 mM phenylmethylsulfonyl fluoride) and recentrifuged as above. The cells were resuspended in 1 mL of TEP buffer and sonicated for three ten-second bursts. The sonicates were made to 1% in Triton X-100, allowed to stand in ice for 5 min, and centrifuged at 8,000xg for 1 min in an Eppendorf Micro Centrifuge (Brinkmann Instruments) to remove the cellular debris. The supernatant was poured into a fresh tube and used for enzyme assays and SDS-PAGE analysis.

A 1 μ L aliquot of the extract of *E. coli* cells containing plasmid pNS2 was assayed for stearyl-ACP desaturase activity in a 5 min reaction, as described above. The extract showed activity of 288 pmol of stearyl-ACP desaturated per min per mL of the extract [The blank (no desaturase enzyme) activity was 15 pmol/min/mL].

Proteins in the extract of *E. coli* cells harboring plasmids pNS2 were resolved by SDS-PAGE, transferred onto Immobilon®-P (Millipore) and cross-reacted with mouse antibody made against purified soybean stearyl-ACP desaturase, as described in Sambrook et al. [(1989) Molecular Cloning: A Laboratory

Manual, 2nd Ed. Cold Spring Harbor Laboratory Press]. The resultant Western blot showed that pNS2 encodes for ca. 42 kD β -galactosidase-stearoyl-ACP desaturase fusion polypeptide.

EXAMPLE 3

5 USE OF SOYBEAN SEED STEAROYL-ACP DESATURASE SEQUENCE IN PLASMID pDS1 AS A RESTRICTION FRAGMENT LENGTH POLYMORPHISM (RFLP) MARKER

Plasmid pDS1 was linearized by digestion with restriction enzyme Eco RI in standard conditions as described in Sambrook et al. [(1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory Press] and labeled with ^{32}P using a Random Priming Kit from Bethesda Research Laboratories Laboratory Press] and labeled with ^{32}P using a Random Priming Kit from Bethesda Research Laboratories Laboratory Press] containing genomic DNA from soybean [*Glycine max* (cultivar Bonus) and *Glycine soja* - (PI81762)], digested with one of several restriction enzymes. After hybridization and washes under standard conditions [Sambrook et al., (1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory Press] autoradiograms were obtained and different patterns of hybridization (polymorphisms) were identified in digests performed with restriction enzymes Pst I and Eco RI. The same probe was then used to map the polymorphic pDS1 loci on the soybean genome, essentially as described by Helentjaris et al. [(1986) Theor. Appl. Genet. 72:761-769]. Plasmid pDS1 probe was applied, as described above, to Southern blots of Eco RI or Pst I digested genomic DNAs isolated from 68 F2 progeny plants resulting from a *G. max* Bonus x *G. soja* PI81762 cross. The bands on the autoradiograms were interpreted as resulting from the inheritance of either paternal (Bonus) or maternal (PI81762) pattern, or both (a heterozygote). The resulting data were subjected to genetic analysis using the computer program Mapmaker [Lander et al., (1987) Genomics 1: 174-181]. In conjunction with previously obtained data for 436 anonymous RFLP markers in soybean [Tingey et al. (1990) J. Cell. Biochem., Supplement 14E p. 291, abstract R153], we were able to position four genetic loci corresponding to the pDS1 probe on the soybean genetic map. This information will be useful in soybean breeding targeted towards developing lines with altered saturate levels, especially for the high stearic acid mutant phenotype, since these recessive traits are most likely be due to loss of seed stearoyl-ACP desaturase enzyme.

35

40

45

50

55

SEQUENCE LISTING

5 (1) GENERAL INFORMATION:

(i) APPLICANT: Hitz, William D.
Yadav, Narendra S

10

(ii) TITLE OF THE INVENTION: Nucleotide
Sequence of Soybean Stearoyl-ACP
Desaturase cDNA

15

(iii) NUMBER OF SEQUENCES: 5

20

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: E. I. du Pont de
Nemours and Company

25

(B) STREET: 1007 Market Street

(C) CITY: Wilmington

(D) STATE: Delaware

30

(E) COUNTRY: USA

(F) ZIP: 19898

35

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: DISKETTE, 3.50
inch, 1.0 MB

40

(B) COMPUTER: Apple Macintosh

(C) OPERATING SYSTEM:

(D) SOFTWARE:

45

(vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER: 07/529,049

(B) FILING DATE: 25-MAY-1990

50

(C) CLASSIFICATION:

55

(vii) ATTORNEY/AGENT INFORMATION;

- (A) NAME: Bruce W. Morrissey
- (B) REGISTRATION NUMBER: 30,663
- (C) REFERENCE/DOCKET NUMBER: BB-1022

(viii) TELECOMMUNICATION INFORMATION:

- (A) TELEPHONE: (302) 892-4927
- (B) TELEFAX: (302) 892-7949
- (C) TELEX: 835420

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 2243 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA to mRNA

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: No

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Glycine max
- (B) STRAIN: Cultivar Wye
- (D) DEVELOPMENTAL STAGE: Developing seeds

(vii) IMMEDIATE SOURCE:

- (A) LIBRARY: cDNA to mRNA
- (B) CLONE: pDS1

(ix) FEATURE:

(A) NAME/KEY:

- 5 (i) 5' non-coding sequence
- (ii) Putative translation
initiation codon
- 10 (iii) Putative transit
peptide coding sequence
- (iv) Mature protein coding
sequence
- 15 (v) Translation termination
codon
- (vi) 3' non-coding sequence

20 (B) LOCATION:

- (i) nucleotides 1 through 69
- (ii) nucleotides 70 through 72
- 25 (iii) nucleotides 70 through 165
- (iv) nucleotides 166 through
1242
- 30 (v) nucleotides 1243 through
1245
- (vi) nucleotides 1246 through
2243

35 (C) IDENTIFICATION METHOD:

- (i) deduced by proximity to
ii) below
- 40 (ii) similarity of the context
of the methionine codon in
the open reading frame to
translation initiation
45 codons of other plastid
transit peptides
- (iii) deduced by proximity to
50 ii) above and iv) below

55

- 5 (iv) experimental determination
of N-terminal amino acid
sequence and subunit size
of purified soybean seed
stearoyl-ACP desaturase
- 10 (v) The translation
termination codon ends
the open reading frame for
a protein of the expected
size
- 15 (vi) established by proximity
to v) above

20 (D) OTHER INFORMATION:
Extracts of E. coli expressing the
mature protein as a fusion protein
show stearoyl-ACP desaturase
25 activity and produce a protein
that cross-reacts to stearoyl-ACP
desaturase antibody

30 (x) PUBLICATION INFORMATION: Sequence not
published.

35 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

CTTCTACATT ACTCTCTCTT CTCCTAAAAA TTTCTAATGC 40

40 TTCCATTGCT TCATCTGACT CACTCATCA ATG GCT CTG AGA CTG AAC CCT 90
Met Ala Leu Arg Leu Asn Pro
-32 -30

45 ATC CCC ACC CAA ACC TTC TCC CTC CCC CAA ATG CCC AGC CTC AGA 135
Ile Pro Thr Gln Thr Phe Ser Leu Pro Gln Met Pro Ser Leu Arg
-25 -20 -15

50 TCT CCC CGC TTC CGC ATG GCT TCC ACC CTC CGC TCC GGT TCC AAA 180
Ser Pro Arg Phe Arg Met Ala Ser Thr Leu Arg Ser Gly Ser Lys
-10 -5 1 5

55

	GAG GTT GAA AAT ATT AAG AAG CCA TTC ACT CCT CCC AGA GAA GTG	225
	Glu Val Glu Asn Ile Lys Lys Pro Phe Thr Pro Pro Arg Glu Val	
	10 15 20	
5	CAT GTT CAA GTA ACC CAC TCT ATG CCT CCC CAG AAG ATT GAG ATT	270
	His Val Gln Val Thr His Ser Met Pro Pro Gln Lys Ile Glu Ile	
	25 30 35	
10	TTC AAA TCT TTG GAG GAT TGG GCT GAC CAG AAC ATC TTG ACT CAT	315
	Phe Lys Ser Leu Glu Asp Trp Ala Asp Gln Asn Ile Leu Thr His	
	40 45 50	
15	CTT AAA CCT GTA GAA AAA TGT TGG CAA CCA CAG GAT TTT TTA CCC	360
	Leu Lys Pro Val Glu Lys Cys Trp Gln Pro Gln Asp Phe Leu Pro	
	55 60 65	
	GAC CCC TCC TCA GAT GGA TTT GAA GAG CAA GTG AAG GAA CTG AGA	405
	Asp Pro Ser Ser Asp Gly Phe Glu Glu Gln Val Lys Glu Leu Arg	
	70 75 80	
20	GAG AGA GCA AAG GAG ATT CCA GAT GAT TAC TTT GTT GTT CTT GTC	450
	Glu Arg Ala Lys Glu Ile Pro Asp Asp Tyr Phe Val Val Leu Val	
	85 90 95	
25	GGA GAC ATG ATC ACA GAG GAA GCT CTG CCT ACT TAC CAA ACT ATG	495
	Gly Asp Met Ile Thr Glu Glu Ala Leu Pro Thr Tyr Gln Thr Met	
	95 100 110	
30	TTA AAT ACT TTG GAT GGA GTT CGT GAT GAA ACA GGT GCC AGC CTT	540
	Leu Asn Thr Leu Asp Gly Val Arg Asp Glu Thr Gly Ala Ser Leu	
	115 120 125	
	ACT TCC TGG GCA ATT TGG ACA AGG GCA TGG ACT GCT GAA GAA AAC	585
	Thr Ser Trp Ala Ile Trp Thr Arg Ala Trp Thr Ala Glu Glu Asn	
	130 135 140	
35	AGA CAC GGT GAT CTT CTT AAC AAA TAT CTG TAC TTG AGT GGA CGA	630
	Arg His Gly Asp Leu Leu Asn Lys Tyr Leu Tyr Leu Ser Gly Arg	
	145 150 155	
40	GTT GAC ATG AAA CAA ATT GAG AAG ACA ATT CAG TAC CTT ATT GGG	675
	Val Asp Met Lys Gln Ile Glu Lys Thr Ile Gln Tyr Leu Ile Gly	
	160 165 170	
	TCT GGG ATG GAT CCT CGG ACC GAG AAC AGC CCC TAC CTT GGT TTC	720
	Ser Gly Met Asp Pro Arg Thr Glu Asn Ser Pro Tyr Leu Gly Phe	
	175 180 185	
45	ATT TAC ACT TCA TTT CAA GAG AGG GCA ACC TTC ATA TCC CAC GGA	765
	Ile Tyr Thr Ser Phe Gln Glu Arg Ala Thr Phe Ile Ser His Gly	
	190 195 200	
50	AAC ACG GCC AGG CTT GCG AAG GAG CAT GGT GAC ATA AAA TTG GCA	810
	Asn Thr Ala Arg Leu Ala Lys Glu His Gly Asp Ile Lys Leu Ala	
	205 210 215	

CAG ATC TGC GGC ATG ATT GCC TCA GAT GAG AAG CGC CAC GAG ACT 855
 Gln Ile Cys Gly Met Ile Ala Ser Asp Glu Lys Arg His Glu Thr 230
 220

5 GCA TAC ACA AAG ATA GTG GAA AAG CTG TTT GAG GTT GAT CCT GAT 900
 Ala Tyr Thr Lys Ile Val Glu Lys Leu Phe Glu Val Asp Pro Asp 245
 235 240

10 GGT ACA GTT ATG GCA TTT GCC GAC ATG ATG AGG AAG AAG ATT GCT 945
 Gly Thr Val Met Ala Phe Ala Asp Met Met Arg Lys Lys Ile Ala 260
 250 255

15 ATG CCA GCA CAC CTT ATG TAT GAC GGC CGC GAC GAC AAC CTG TTT 990
 Met Pro Ala His Leu Met Tyr Asp Gly Arg Asp Asp Asn Leu Phe 275
 265 270

20 GAT AAC TAC TCT GCC GTC GCG CAG CGC ATT GGG GTC TAC ACT GCA 1035
 Asp Asn Tyr Ser Ala Val Ala Gln Arg Ile Gly Val Tyr Thr Ala 290
 280 285

25 AAG GAC TAT GCT GAC ATA CTC GAA TTT CTG GTG GGG AGG TGG AAG 1080
 Lys Asp Tyr Ala Asp Ile Leu Glu Phe Leu Val Gly Arg Trp Lys 305
 295 300

30 GTG GAG CAG CTA ACC GGA CTT TCA GGT GAG GGA AGA AAG GCT CAG 1125
 Val Glu Gln Leu Thr Gly Leu Ser Gly Glu Gly Arg Lys Ala Gln 320
 310 315

35 GAA TAC GTT TGT GGG CTG CCA CCA AGA ATC AGA AGG TTG GAG GAG 1170
 Glu Tyr Val Cys Gly Leu Pro Pro Arg Ile Arg Arg Leu Glu Glu 335
 325 330

40 AGA GCT CAA GCA AGA GGC AAG GAG TCG TCA ACA CTT AAA TTC AGT 1215
 Arg Ala Gln Ala Arg Gly Lys Glu Ser Ser Thr Leu Lys Phe Ser 350
 340 345

45 TGG ATT CAT GAC AGG GAA GTA CTA CTC TAAATGCT TGCACCAAGG 1260
 Trp Ile His Asp Arg Glu Val Leu Leu 359
 355

50 GAGGAGCATG GTGAATCTTC CAGCAATACC ATTCTGAGAA ATGTTGAATA 1310
 GTTGAAAATT CAGTTTGTCA TTTTATCTT TTTTCTCC TGTTTTTGG 1360
 TCTTATGTTA TATGCCACTG TAAGGTGAAA CAGTTGTTCT TGCATGGTTC 1410
 GCAAGTTAAG CAGTTAGGGG CAGCTGTAGT ATTAGAAATG CTATTTTTTG 1460
 TTTCCCTTTT CTGTGGTAGT GATGTCTGTG GAAGTATAAG TAAACGTTTT 1510
 TTTTCTC TGGCAATTTTG ATGATAAAGA AAATTTAGTT CTAAAAACCG 1560
 TCGCACCTTC CCTGAGGCTT CTCTGTCTG TCGCGAGTGA CCATGGTGAG 1610
 GGTTAGTGTG CTGAACGATG CTCTGAAGAG CATGTACAAT GCTGAGAAAA 1660
 GGGGAAAGCG CCAAGTCATG ATTCGGCCAT CCTCCAAAGT CATTATCAAA 1710

TTCCTTTTGG TGATGCAGAA GCACGGATAC ATTGGAGAGT TTGAGTATGT 1760
 TGATGACCAC AGGGCTGGTA AAATCGTGGT TGAATTGAAC GGTAGACTGA 1810
 5 ACAAGTGTGG GGTATTAGT CCCCCTTTTG ATGTCGGCGT CAAAGAGATT 1860
 GAAGGTTGGA CTGCTAGGCT TCTCCCTCA AGACAGTTTG GGTATATTGT 1910
 ATTGACTACC TCTGCCGCA TCATGGATCA CGAAGAAGCT AGGAGAAAAA 1960
 10 ATGTTGGTGG TAAGGTACTG GGTTCCTCT ACTAGAGTTT AATTCGATT 2010
 AAGAGGATGT CAGGAATTC AATTGAGATT CATGGATTGT AATGGAGGAT 2060
 ATGCTAGGCC CCTAGTAATA TCAAGCATAG CAGGAGCTGT TTTGTGATGT 2110
 15 TCCTTATTTT GTTTGCAAAA CCAAGTTGGT AACTATAACT TTTATTTTCT 2160
 TTTATCATTA TTTTCTTTA TACCAAATG TACTGGCCAA GTTGTTTTAA 2210
 20 ACAGTGAGAA CTTTGATTAG AAAAAAAAAA AAA 2243

(2) INFORMATION FOR SEQ ID NO:2:

25

(i) SEQUENCE CHARACTERISTICS:

30

- (A) LENGTH: 216 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

35

(ii) MOLECULE TYPE: cDNA to mRNA

40

(iii) HYPOTHETICAL: No

(iv) ANTISENSE: No

45

(vi) ORIGINAL SOURCE:

50

- (A) ORGANISM: Glycine max
- (B) STRAIN: Cultivar Wye
- (D) DEVELOPMENTAL STAGE: Developing seeds

55

(vii) IMMEDIATE SOURCE:

(A) LIBRARY: cDNA to mRNA

(B) CLONE: pDS4a

(ix) FEATURE:

(A) NAME/KEY:3' non-coding sequence

(B) LOCATION: nucleotides 1 through
216

(C) IDENTIFICATION METHOD: Homology of
clones pDS4a and pDS1
and similarity of
sequence in SEQ ID NO:1
to 3' non-coding
sequence in SEQ ID NO:1

(x) PUBLICATION INFORMATION: Sequence not
published.

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

GAAATGTTGA ATAGTTGAAA ATTCAGTTTG TCATTTTAT CTTTTATTTT 50
TTCTCCTTTT TTGGTCTTTG TTATATGTCA CTGTAAGGTG AAGCAGTTGT 100
TCTTGCATGG TTCGCAAGTT AAGCAGTTAG GGGCAGCTGT AGTATTAGAA 150
ATGGTATTTT TTTTTTGTG TTCGCTTTTC TCTGTGGTAG TGATGTCTGT 200
CGAAGTATAA GTAAAC 216

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 amino acids

(B) TYPE: amino acid

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(iii) HYPOTHETICAL: No

5 (v) FRAGMENT TYPE: N-terminal fragment

(vi) ORIGINAL SOURCE:

10 (A) ORGANISM: Glycine max
(B) STRAIN: Cultivar Wye
(C) DEVELOPMENTAL STAGE: Developing
15 seeds

(ix) FEATURE:

20 (A) NAME/KEY: N-terminal sequence
(B) LOCATION: 1 through 16 amino acid
residues
(C) IDENTIFICATION METHOD: N-terminal
25 amino acid sequencing

(x) PUBLICATION INFORMATION: Sequence not
30 published

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

35 Arg Ser Gly Ser Lys Glu Val Glu Asn Ile Lys Lys Pro Phe Thr Pro
1 5 10 15

40 (2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:

45 (A) LENGTH:36 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
50

55

(ii) MOLECULE TYPE: Other nucleic acid: mixture
of oligonucleotides

5

(iii) HYPOTHETICAL: Yes

(ix) FEATURE:

10

(A) NAME/KEY: Coding sequence

(B) LOCATION: 1 through 36 bases

15

(x) PUBLICATION INFORMATION : Sequence not
published

20

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

AAR	GAR	GTN	GAR	AAY	ATH	AAR	AAR	CCN	TTY	ACN	CCN	3
Lys	Glu	Val	Glu	Asn	Ile	Lys	Lys	Pro	Phe	Thr	Pro	
1				5					10			

25

(2) INFORMATION FOR SEQ ID NO:5:

30

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH:35 base pairs

(B) TYPE: nucleic acid

35

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

40

(ii) MOLECULE TYPE: Other nucleic acid: mixture
of synthetic oligonucleotides

(ix) FEATURE:

45

(C) OTHER INFORMATION: N at positions
3,6,9, and 27 is deoxyinosine.

50

(x) PUBLICATION INFORMATION: Sequence not
published

55

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

5 GGNGTNAANG GCTTCTTRAT RTTYTCNACN TCCTT 35

10 Claims

1. An isolated nucleic acid fragment comprising a nucleotide sequence encoding the soybean seed stearyl-ACP desaturase corresponding to the nucleotides 1 to 2243 in SEQ ID NO:1, or any soybean nucleic acid fragment substantially homologous therewith encoding a functional stearyl-ACP desaturase.
2. An isolated nucleic acid fragment of Claim 1 wherein said nucleotide sequence encodes the soybean seed stearyl-ACP desaturase precursor corresponding to nucleotides 70-1245 in SEQ ID NO:1, or any soybean nucleic acid fragment substantially homologous therewith encoding a functional stearyl-ACP desaturase precursor.
3. A nucleic acid fragment of Claim 2, wherein the said nucleotide sequence encodes the mature soybean seed stearyl-ACP desaturase enzyme, corresponding to nucleotides 166 to 1245 in SEQ ID NO:1.
4. A chimeric gene capable of transforming a soybean plant cell comprising a nucleic acid fragment of Claim 1 operably linked to suitable regulatory sequences producing antisense inhibition of soybean seed stearyl-ACP desaturase in the seed.
5. A chimeric gene capable of transforming a plant cell of an oil-producing species comprising a nucleic acid fragment of Claim 2 operably linked to suitable regulatory sequences resulting in overexpression of said soybean seed stearyl-ACP desaturase in the plastid of said plant cell.
6. A chimeric gene capable of transforming a plant cell of an oil-producing species comprising a nucleic acid fragment of Claim 3 operably linked to suitable regulatory sequences resulting in the expression of said mature soybean seed stearyl-ACP desaturase enzyme in the cytoplasm of said plant cell.
7. A method of producing soybean seed oil containing higher-than-normal levels of stearic acid comprising:
 - (a) transforming a soybean plant cell with a chimeric gene of Claim 4,
 - (b) growing fertile soybean plants from said transformed soybean plant cells,
 - (c) screening progeny seeds from said fertile soybean plants for the desired levels of stearic acid, and
 - (d) crushing said progeny seed to obtain said soybean oil containing higher-than-normal levels of stearic acid.
8. A method of producing oils from plant seed containing lower-than-normal levels of stearic acid comprising:
 - (a) transforming a plant cell of an oil producing species with a chimeric gene of Claims 5 or 6,
 - (b) growing sexually mature plants from said transformed plant cells of an oil producing species,
 - (c) screening progeny seeds from said fertile plants for the desired levels of stearic acid, and
 - (d) crushing said progeny seed to obtain said oil containing lower-than-normal levels of stearic acid.
9. A method of Claim 8 wherein said plant cell of an oil producing species is selected from the group consisting of soybean, rapeseed, sunflower, cotton, cocoa, peanut, safflower, and corn.
10. A method of Claim 7 wherein said step of transforming is accomplished by a process selected from the group consisting of Agrobacterium infection, electroporation, and high-velocity ballistic bombardment.

11. A method of Claim 8 wherein said step of transforming is accomplished by a process selected from the group consisting of Agrobacterium infection, electroporation, and high-velocity ballistic bombardment.
12. A method of producing mature soybean seed stearoyl-ACP desaturase enzyme in microorganisms comprising:
- (a) transforming a microorganism with a chimeric gene of Claim 6,
 - (b) growing said transformed microorganism to produce quantities of said mature soybean seed stearoyl-ACP desaturase enzyme, and
 - (c) isolating and purifying said mature soybean seed stearoyl-ACP desaturase enzyme.
13. A method of breeding soybean plants producing altered stearic acid levels in seed oil due to altered levels of stearyl-ACP desaturase in said soybean plants by RFLP mapping comprising:
- (a) making a cross between two soybean varieties differing in stearic acid levels due to altered levels of stearyl-ACP desaturase;
 - (b) making a Southern blot of genomic DNA isolated from several progeny plants resulting from the cross following digestion with a suitable restriction enzyme that reveals polymorphism linked to the altered levels of stearic acid using a radiolabelled nucleic acid fragment of Claim 1 as a hybridization probe;
 - (c) hybridizing the Southern blot with the radiolabelled nucleic acid fragment of Claim 1; and
 - (d) selecting said soybean plants that inherit the RFLP linked to the desired level of stearic acid.

Patentansprüche

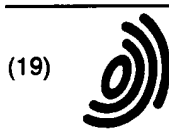
- Isoliertes Nukleinsäurefragment, umfassend eine Nukleotidsequenz, die für Sojabohnensamen-Stearoyl-ACP-Desaturase kodiert, die den Nukleotiden 1 - 2243 in SEQ ID NO:1 entspricht, oder ein Sojabohnen-Nukleinsäurefragment, das im wesentlichen dazu homolog ist, das für eine funktionelle Stearoyl-ACP-Desaturase kodiert.
- Isoliertes Nukleinsäurefragment nach Anspruch 1, worin die genannte Nukleotidsequenz für die Sojabohnensamen-Stearoyl-ACP-Desaturase-Vorstufe, entsprechend den Nukleotiden 70 - 1245 in SEQ ID NO:1, kodiert, oder ein Sojabohnen-Nukleinsäurefragment, das im wesentlichen dazu homolog ist und für eine funktionelle Stearoyl-ACP-Desaturase-Vorstufe kodiert.
- Nukleinsäurefragment nach Anspruch 2, bei dem die genannte Nukleotidsequenz für das Stearoyl-ACP-Desaturase-Enzym von reifem Sojabohnensamen kodiert, das den Nukleotiden 166 - 1245 in SEQ ID NO:1 entspricht.
- Chimäres Gen, das in der Lage ist, eine Sojabohnen-Pflanzenzelle zu transformieren, umfassend ein Nukleinsäurefragment nach Anspruch 1, das zweckorientiert mit geeigneten regulatorischen Sequenzen verknüpft ist, die eine Antisinn-Hemmung der Sojabohnensamen-Stearoyl-ACP-Desaturase in dem Samen erzeugen.
- Chimäres Gen, das in der Lage ist, eine Pflanzenzelle einer ölproduzierenden Spezies zu transformieren, umfassend ein Nukleinsäurefragment nach Anspruch 2, das mit geeigneten regulatorischen Sequenzen zweckorientiert verknüpft ist, was zu einer Überexpression der genannten Sojabohnensamen-Stearoyl-ACP-Desaturase in dem Plastid der genannten Pflanzenzelle führt.
- Chimäres Gen, das in der Lage ist, eine Pflanzenzelle einer ölproduzierenden Spezies zu transformieren, umfassend ein Nukleinsäurefragment nach Anspruch 3, das mit geeigneten regulatorischen Sequenzen zweckorientiert verknüpft ist, was zu der Expression des genannten Stearoyl-ACP-Desaturase-Enzyms von reifem Sojabohnensamen in dem Cytoplasma der genannten Pflanzenzelle führt.
- Verfahren zur Herstellung von Sojabohnensamenöl, enthaltend höhere als normale Konzentrationen an Stearinsäure, umfassend:
 - Transformieren einer Sojabohnen-Pflanzenzelle mit einem chimären Gen nach Anspruch 4,
 - Züchten der fruchtbaren Sojabohnenpflanzen aus den genannten transformierten Sojabohnen-Pflanzenzellen,

- (c) Überprüfung der zeugungsfähigen Samen aus den genannten fruchtbaren Sojabohnenpflanzen auf die gewünschten Stearinsäure-Konzentrationen und
 (d) Zerquetschen des genannten zeugungsfähigen Samens, um das genannte Sojabohnenöl zu erhalten, das höhere als normale Stearinsäure-Konzentrationen enthält.
- 5 8. Verfahren zur Herstellung von Ölen aus Pflanzensamen, die niedrigere als normale Stearinsäure-Konzentrationen enthalten, umfassend:
 (a) Transformieren einer Pflanzenzelle einer ölproduzierenden Spezies mit einem chimären Gen nach den Ansprüchen 5 oder 6,
 10 (b) Züchten von sexuell reifen Pflanzen aus den genannten transformierten Pflanzenzellen einer ölproduzierenden Spezies,
 (c) Überprüfung der zeugungsfähigen Samen aus den genannten fruchtbaren Pflanzen auf die gewünschten Stearinsäure-Konzentrationen, und
 (d) Zerquetschen des genannten zeugungsfähigen Samens, um das genannte Öl zu erhalten, das
 15 niedrigere als normale Stearinsäure-Konzentrationen enthält.
9. Verfahren nach Anspruch 8, bei dem die genannte Pflanzenzelle einer ölproduzierenden Spezies aus der Gruppe ausgewählt wird, bestehend aus Sojabohne, Rapssamen, Sonnenblume, Baumwolle, Kakao, Erdnuß, Färberdistel und Mais.
- 20 10. Verfahren nach Anspruch 7, bei dem die genannte Stufe der Transformation durch ein Verfahren durchgeführt wird, ausgewählt aus der Gruppe, bestehend aus einer Agrobacterium-Infektion, einer Elektroporation und einer Hochgeschwindigkeitsstoßbombardierung.
- 25 11. Verfahren nach Anspruch 8, bei dem die genannte Transformationsstufe durch ein Verfahren durchgeführt wird, ausgewählt aus der Gruppe, bestehend aus einer Agrobacterium-Infektion, einer Elektroporation und einer Hochgeschwindigkeitsstoßbombardierung.
- 30 12. Verfahren zur Herstellung des Stearoyl-ACP-Desaturase-Enzyms von reifem Sojabohnensamen in Mikroorganismen, umfassend:
 (a) Transformieren eines Mikroorganismus mit einem chimären Gen nach Anspruch 6,
 (b) Züchten des genannten transformierten Mikroorganismus, um Mengen des genannten Stearoyl-ACP-Desaturase-Enzyms von reifem Sojabohnensamen zu produzieren und
 (c) Isolieren und Reinigen des genannten Stearoyl-ACP-Desaturase-Enzyms von reifem Sojabohnensamen.
 35
13. Verfahren zur Züchtung von Sojabohnenpflanzen, die aufgrund veränderter Konzentrationen der Stearoyl-ACP-Desaturase in den genannten Sojabohnenpflanzen durch RFLP-Kartierung veränderte Stearinsäure-Konzentrationen in dem Samenöl produzieren, umfassend:
 40 (a) Kreuzen zweier Sojabohnen-Varietäten, die sich aufgrund der veränderten Konzentrationen der Stearoyl-ACP-Desaturase in den Stearinsäure-Konzentrationen unterscheiden,
 (b) Anfertigen eines Southern Blots der genomischen DNA, die aus mehreren, aus der Kreuzung hervorgegangenen zeugungsfähigen Pflanzen isoliert worden ist, und anschließender Verdau mit einem geeigneten Restriktionsenzym, das den Polymorphismus, der mit den veränderten Stearinsäure-Konzentrationen verknüpft ist, aufdeckt, wobei ein radioaktiv markiertes Nukleinsäurefragment
 45 nach Anspruch 1 als Hybridisierungssonde verwendet wird,
 (c) Hybridisierung des Southern Blots mit dem radioaktiv markierten Nukleinsäurefragment nach Anspruch 1, und
 (d) Auswählen der genannten Sojabohnenpflanzen, die das RFLP, das mit der gewünschten
 50 Stearinsäure-Konzentration verknüpft ist, vererben.

Revendications

- 55 1. Un fragment d'acide nucléique isolé comprenant une séquence nucléotidique codant pour la stéaroyl-ACP-désaturase de graine de soja correspondant aux nucléotides 1 à 2243 de SEQ ID N° 1, ou tout fragment d'acide nucléique de soja sensiblement homologue à celui-ci codant pour une stéaroyl-ACP-désaturase fonctionnelle.

2. Un fragment d'acide nucléique isolé de la revendication 1, dans lequel ladite séquence nucléotidique code pour le précurseur de stéaroyl-ACP-désaturase de graine de soja correspondant aux nucléotides 70 à 1245 de SEQ ID N° 1, ou tout fragment d'acide nucléique de soja sensiblement homologue à celui-ci codant pour un précurseur de stéaroyl-ACP-désaturase fonctionnel.
3. Un fragment d'acide nucléique de la revendication 2, dans lequel ladite séquence nucléotidique code pour la stéaroyl-ACP-désaturase de graine de soja mûre, correspondant aux nucléotides 166 à 1245 de SEQ ID N° 1.
4. Un gène chimérique capable de transformer une cellule de soja, comprenant un fragment d'acide nucléique de la revendication 1, lié fonctionnellement à des séquences régulatrices appropriées produisant une inhibition anti-sens de la stéaroyl-ACP-désaturase dans la graine.
5. Un gène chimérique capable de transformer une cellule végétale d'une espèce productrice d'huile, comprenant un fragment d'acide nucléique de la revendication 2 lié fonctionnellement à des séquences régulatrices appropriées donnant lieu à une surexpression de ladite stéaroyl-ACP-désaturase de graine de soja dans le plastide de ladite cellule végétale.
6. Un gène chimérique capable de transformer une cellule végétale d'une espèce productrice d'huile, comprenant un fragment d'acide nucléique de la revendication 3 lié fonctionnellement à des séquences régulatrices appropriées donnant lieu à l'expression de ladite stéaroyl-ACP-désaturase de graine de soja mûre dans le cytoplasme de ladite cellule végétale.
7. Un procédé de production d'huile de graine de soja contenant des taux supérieurs à la normale d'acide stéarique, consistant à :
 - (a) transformer une cellule de soja avec un gène chimérique de la revendication 4,
 - (b) faire croître des plants de soja fertiles à partir de cellules de soja transformées,
 - (c) sélectionner des graines de descendance provenant desdits plants de soja fertiles pour les taux souhaités d'acide stéarique, et
 - (d) broyer lesdites graines de descendance pour obtenir ladite huile de soja contenant des taux d'acide stéarique supérieurs à la normale.
8. Un procédé de production d'huiles à partir de graines végétales contenant des taux d'acide stéarique inférieurs à la normale, consistant à :
 - (a) transformer une cellule végétale d'une espèce productrice d'huile avec un gène chimérique de la revendication 5 ou 6,
 - (b) faire croître des plants sexuellement matures à partir desdites cellules végétales transformées d'une espèce productrice d'huile,
 - (c) sélectionner des graines de descendance provenant desdits plants fertiles pour les taux désirés d'acide stéarique, et
 - (d) broyer lesdites graines de descendance pour obtenir l'huile contenant des taux d'acide stéarique inférieurs à la normale.
9. Un procédé de la revendication 8, dans lequel ladite cellule végétale d'une espèce productrice d'huile est choisie dans le groupe formé par le soja, le colza, le tournesol, le cotonnier, le cacaoyer, l'arachide, le carthame et le maïs.
10. Un procédé de la revendication 7, dans lequel ladite étape de transformation est exécutée par un procédé choisi dans le groupe formé par une infection par *Agrobacterium*, une électroporation et un bombardement balistique à grande vitesse.
11. Un procédé de la revendication 8, dans lequel ladite étape de transformation est exécutée par un procédé choisi dans le groupe formé par une infection par *Agrobacterium*, une électroporation et un bombardement balistique à grande vitesse.
12. Un procédé de production de stéaroyl-ACP-désaturase de graine de soja mûre dans des microorganismes, consistant à :
 - (a) transformer un microorganisme avec un gène chimérique de la revendication 6,



Europäisches Patentamt
European Patent Office
Office européen des brevets



(11)

EP 0 804 618 B1

(12)

EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention
of the grant of the patent:
27.01.1999 Bulletin 1999/04

(51) Int Cl.⁶: **C12Q 1/68**

(86) International application number:
PCT/US95/15150

(21) Application number: **95942451.6**

(87) International publication number:
WO 96/17082 (06.06.1996 Gazette 1996/26)

(22) Date of filing: **21.11.1995**

(54) COMPOUND MICROSATELLITE PRIMERS FOR THE DETECTION OF GENETIC POLYMORPHISMS

MIKROSATELLITEVERBINDUNG FÜR DETEKTION GENETISCHES POLYMORPHISMEN

SONDES DE MICROSATELLITES COMPOSES POUR LA DETECTION DE POLYMORPHISMES GENETIQUES

(84) Designated Contracting States:
BE CH DE FR GB IT LI

(30) Priority: **28.11.1994 US 346456**

(43) Date of publication of application:
05.11.1997 Bulletin 1997/45

(73) Proprietor: **E.I. DU PONT DE NEMOURS AND COMPANY**
Wilmington Delaware 19898 (US)

(72) Inventors:
• **MORGANTE, Michele**
I-33100 Udine (IT)
• **VOGEL, Julie, Marie**
Malvern, PA 19355-8717 (US)

(74) Representative: **Cockbain, Jullan, Dr.**
Frank B. Dehn & Co.,
European Patent Attorneys,
179 Queen Victoria Street
London EC4V 4EL (GB)

(56) References cited:
EP-A- 0 534 858 EP-A- 0 552 545
WO-A-91/18114 WO-A-92/13969
WO-A-93/07166 WO-A-94/17203
WO-A-95/15400

- **GENOME**, vol. 36, no. 5, October 1993, OTTAWA, CA, pages 884-889, XP000569548 M. LYNN ET AL.: "mapping maize microsatellites and polymerase chain reaction confirmation of the target repeats using a CT primer"

- **GENOMICS**, vol. 20, no. 2, 15 March 1994, ACADEMIC PRESS INC., NY, US, pages 176-183, XP000569780 E. ZIETKIEWICZ ET AL.: "Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction" cited in the application
- **NUCLEIC ACIDS RESEARCH**, vol. 22, no. 15, 11 August 1994, IRL PRESS LIMITED, OXFORD, ENGLAND, pages 3257-3258, XP002002453 K.-S. WU ET AL.: "Detection of microsatellite polymorphisms without cloning" cited in the application
- **THEORETICAL AND APPLIED GENETICS**, no. 88, 1901 - April 1994, SPRINGER INTERNATIONAL, NY, US, pages 1-6, XP000569790 Z. WANG ET AL.: "Survey of plant short tandem repeats" cited in the application
- **AM.J. HUMAN. GENET.**, vol. 44, 1989, AM.SOC.HUM.GENET., CHICAGO, US, pages 388-396, XP000431459 J.L.WEBER AND P.E. MAY: "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction"
- **BIOTECHNIQUES**, vol. 15, no. 2, 1 August 1993, pages 304-309, XP000382830 GRIST S A ET AL.: "DINUCLEOTIDE REPEAT POLYMORPHISMS ISOLATED BY THE POLYMERASE CHAIN REACTION"
- **PROC. NATL.ACAD SCI.**, vol. 91, no. 12, June 1994, NATL. ACAD SCI., WASHINGTON, DC, US, pages 5466-5470, XP002002454 M.A. S. MAROOF ET AL.: "Extraordinarily polymorphic microsatellite DNA in barley: Species diversity, chromosomal locations, and population dynamics"

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

EP 0 804 618 B1

- PLANT JOURNAL, vol. 3, no. 1, January 1993, BLACKWELL SCIENCE LTD, OXFORD, UK, pages 175-182, XP002002455 M. MORGANTE AND A.M. OLIVERI: "PCR-amplified microsatellites as markers in plant genetics" cited in the application
- NUCLEIC ACIDS RESEARCH, vol. 19, no. 8, 25 April 1991, IRL PRESS LIMITED, OXFORD, ENGLAND, page 1950 XP002002456 M.A.R. YUILLE ET AL.: "Rapid determination of sequences flanking microsatellites"
- NUCLEIC ACIDS RESEARCH, vol. 20, no. 1, 11 January 1992, IRL PRESS LIMITED, OXFORD, ENGLAND, page 141 XP002002457 D.L. BROWNE AND M. LITT: "Characterization of (CA)_n microsatellites with degenerate sequencing primers"

DescriptionFIELD OF INVENTION

5 The present invention relates the use of perfect, compound simple sequence repeats (SSR) as self-anchoring primers for the identification and analysis of DNA sequence polymorphisms. More specifically it has been observed that any one type of simple sequence repeat (SSR) in both plant and animal genomes often exists directly adjacent to an SSR of a different type, usually with perfect periodicity of one of the component nucleotides shared by both SSRs. This observation has allowed the design of self-anchoring primers in new variations of polymerase chain reaction-based multiplexed genome assays, including inter-repeat amplification and amplified fragment length polymorphism assays. These method variations collectively have been termed selective amplification of microsatellite polymorphic loci (SAMPL).

BACKGROUND

15 The ability to map eukaryotic genomes has become an essential tool for the diagnosis of genetic diseases, and for plant breeding and forensic medicine. An absolute requirement for elucidation of any genetic linkage map is the ability to identify DNA sequence variation. The realization that genetic (DNA) polymorphisms between phenotypically identical individuals are present and can be used as markers for genetic mapping has produced major advances in the art of developing eukaryotic linkage maps.

20 Techniques for identifying genetic polymorphisms are relatively few and to date have been time consuming and labor intensive. One of the most common techniques is referred to as restriction fragment length polymorphism or RFLP (Botstein et al. *Am. J. Hum. Genet.* 342, 314, (1980)). Using RFLP technology, genetic markers based on single or multiple point mutations in the genome may be detected by differentiating DNA banding patterns from restriction enzyme analysis. As restriction enzymes cut DNA at specific target site sequences, a point mutation within this site may result in the loss or gain of a recognition site, giving rise in that genomic region to restriction fragments of different length. Mutations caused by the insertion, deletion or inversion of DNA stretches will also lead to a length variation of DNA restriction fragments. Genomic restriction fragments of different lengths between genotypes can be detected with region-specific probes on Southern blots (Southern, E. M., *J. Mol. Biol.* 98, 503, (1975)). The genomic DNA is typically digested with nearly any restriction enzyme of choice. The resulting fragments are electrophoretically size-separated, transferred to a membrane, and then hybridized against a suitably labelled probe for detection of fragments corresponding to a specific region of the genome. RFLP genetic markers are particularly useful in detecting genetic variation in phenotypically silent mutations and serve as highly accurate diagnostic tools. RFLP analysis is a useful tool in the generation of codominant genetic markers but suffers from the need to separate restriction fragments electrophoretically and often requires a great deal of optimization to achieve useful background to signal ratios where significant polymorphic markers can be detected. In addition, the RFLP method relies on DNA polymorphisms existing within actual restriction sites. Any other point mutations in the genome usually go undetected. This is a particularly difficult problem when assaying genomes with inherently low levels of DNA polymorphism. Thus, RFLP differences often are difficult to identify.

40 Another method of identifying polymorphic genetic markers employs DNA amplification using short primers of arbitrary sequence. These primers have been termed 'random amplified polymorphic DNA', or "RAPD" primers, Williams et al., *Nucl. Acids. Res.*, 18, 6531 (1990) and U.S. 5,126,239; (also EP0 543 484 A2, WO 92/07095, WO 92/07948, WO 92/14844, and WO 92/03567). The RAPD method amplifies either double or single stranded nontargeted, arbitrary DNA sequences using standard amplification buffers, dATP, dCTP, dGTP and TTP nucleotides, and a thermostable DNA polymerase such as Taq polymerase. The nucleotide sequence of the primers is typically about 9 to 13 bases in length, between 50 and 80% G+C in composition and contains no palindromic sequences. Differences as small as single nucleotides between genomes can affect the RAPD primer's binding/target site, and a PCR product may be generated from one genome but not from another. RAPD detection of genetic polymorphisms represents an advance over RFLP in that it is less time consuming, more informative, and readily adaptable to automation. The use of the RAPD assay is limited, however, in that only dominant polymorphisms can be detected; this method does not offer the ability to examine simultaneously all the alleles at a locus in a population. Nevertheless, because of its sensitivity for the detection of polymorphisms, RAPD analysis and variations based on RAPD/PCR methods have become the methods of choice for analyzing genetic variation within species or closely related genera, both in the animal and plant kingdoms.

55 A third method more recently introduced for identifying and mapping genetic polymorphisms is termed amplified fragment length polymorphism or AFLP (M. Zabeau, EP 534,858). AFLP is similar in concept to RFLP in that restriction enzymes are used to specifically digest the genomic DNA to be analyzed. The primary difference between these two methods is that the amplified restriction fragments produced in AFLP are modified by the addition of specific, known

adaptor sequences which serve as the target sites for PCR amplification with adaptor-directed primers. Briefly, restriction fragments are generated from genomic DNA by complete digestion with a single or double restriction enzyme combination, the latter using a "frequent" cutter combined with a "rare" cutter. Optimal results are obtained when one of these enzymes has a tetranucleotide recognition site, and the other enzyme a hexanucleotide site. Such a double enzyme digestion generates a mixture of single- and double-digested genomic DNA fragments. Next, double-stranded adaptors composed of synthetic oligonucleotides of moderate length (10-30 bases) are specifically ligated to the ends of the restriction fragments. The individual adaptors corresponding to the different restriction sites all carry distinct DNA sequences.

One of the adaptors, usually the one corresponding to the hexanucleotide-site restriction enzyme, carries a biotin moiety. The application of biotin-streptavidin capture methodology leads to the selective removal of all nonbiotinylated restriction fragments (those bordered at both ends by the tetranucleotide restriction site), and thus effectively enriches the population for fragments carrying the biotinylated adaptor at one or both ends. As a result, the DNA fragment mixture is also enriched for asymmetric fragments, those carrying a different restriction site at each end. The selected fragments serve as templates for PCR amplification using oligonucleotide primers that correspond to the adaptor/restriction site sequences. These adaptor-directed primers also can include at their 3' ends from 1 to 10 arbitrary nucleotides, which will anneal to and prime from the genomic sequence directly adjacent to the restriction site on the DNA fragment.

A PCR reaction using this type of pooled fragment template and adaptor-directed primers results in the co-amplification of multiple genomic fragments. Any DNA sequence differences between genomes in the region of the restriction sites or the 1-10 nucleotides directly adjacent to the restriction sites leads to differences, or polymorphisms (dominant and codominant), in the PCR products generated. Multiple fragments are simultaneously co-amplified, and some proportion of these will be polymorphic between genomes.

A fourth method of assaying polymorphisms has involved utilizing the high degree of length variation resulting from certain repeating nucleotide sequences found in most genomes. Most if not all eukaryotic genomes are populated with repeating base sequences variously termed simple sequence repeats (SSR), simple sequence length polymorphisms (SSLP), dinucleotide, trinucleotide, tetranucleotide, or pentanucleotide repeats, and microsatellites. Simple sequence repeats have been demonstrated to be useful as genetic markers (DE 38 34 636 A, Jackle et al.; Weber, et al, *Am J Hum Genet* 44, 388, (1989); Litt et al., *Am J Hum Genet* 4, 397, (1989)). Weber et al., (*Genomics* 11, 695, (1991)) have successfully used SSRs for comparative analysis and mapping of mammalian genomes, and several groups (Akkaya, et al., *Genetics* 132, 1131 (1992), Morgante & Olivieri, *Plant J* 3, 175 (1993), Wu & Tanksley, *Mol. Gen. Genet.* 241, 225, (1993)) have demonstrated similar results with plant genomes. SSR polymorphisms can be detected by PCR using minute amounts of genomic DNA and, unlike RAPDs, they provide codominant markers and can detect a high degree of genetic polymorphism (Weber, J. L. (1990) In Tilghman, S., Daves, K., (eds) 'Genome Analysis vol. 1: Genetic and physical mapping', Cold Spring Harbor Laboratory Press, pp 159-181.).

Although SSR-directed PCR primers are highly effective for detecting polymorphism, their use suffers from a variety of practical drawbacks. Typically, markers generated by these methods are obtained by first constructing a genomic library, screening the library with probes representing the core elements of a particular repeat sequence, purifying and sequencing the positive clones, and synthesizing the primers specific for the flanking sequences for each cloned SSR locus. Genomic DNA is then amplified to screen for polymorphisms, and mapping of the genome is then carried out. The entire process is time consuming, expensive and technically demanding, and as a result has been somewhat limited in its application.

At least one method has been developed as an attempt to circumvent these limitations and allow the use of polymorphic SSR markers more directly, with no *a priori* knowledge of particular SSR locus sequences. Zietkiewicz, et al., (*Genomics*, 20, 176, (1994)), for example, demonstrate that a single-primer PCR amplification can be used to detect length polymorphisms between adjacent (CA)_n repeats in animal and plant genomes. The PCR primers used for this assay each contain a particular SSR sequence that is flanked immediately 5' or 3' by 2 to 4 nucleotides of known or arbitrary sequence; these anchor sequences anneal to the non-SSR genomic sequences that flank the SSR sequence in the genome and serve to "anchor" the primer to a single position at each matching SSR locus. Radiolabeled SSR-to-SSR amplification products, generated when adjacent SSR sequences are oppositely oriented and spaced closely enough in the genome, are analysed by gel electrophoresis followed by autoradiography. This approach eliminates the need for cloning and sequencing SSRs from the genome, and reveals an enriched polymorphic banding pattern relative to single-locus SSR. Zietkiewicz et al, attribute the enriched pattern to use of the arbitrary sequence anchor, which allows the SSR primer to anneal and prime from many SSR target loci simultaneously.

In a concept similar to Zietkiewicz, Wu et al., (*Nucleic Acids Res.* 22, 3257, (1994)) teach a method for the detection of polymorphisms where genomic DNA is amplified by asymmetric thermally cycled PCR using radiolabeled 5' anchored primers consisting of microsatellite repeats in the presence of RAPD primers of arbitrary sequence. The method of Wu et al. is useful for the generation of genetic markers that incorporate many features of microsatellite repeats. Wu et al. does not disclose the use of compound microsatellite repeat primers for amplification.

Simple sequence repeats may be classified in several ways. In categorizing and characterizing human (CA)_n or (GT)_n repeats, Weber, J. L. (*Genome Analysis* Vol. 1, 159, 1990, Cold Spring harbor Laboratory Press, NY) defines at least three types of (CA)_n SSR: simple perfect SSR, simple imperfect SSR, and compound SSR. Each perfect SSR is considered to be a simple (CA)_n tandem sequence, with no interruptions within the repeat. Imperfect SSR are defined as those repeating sequences with one or more interruptions of up to 3 nonrepeat bases within the run of the repeat. Compound SSR are defined as those sequences with a CA or GT repeat stretch adjacent to or within 3 nucleotides of a block of short tandem repeats of a different sequence. Weber notes that perfect sequence repeats in humans comprise about 65% of the total (dC-dA)_n(dG-dT)_n sequences cloned from the genome, imperfect repeats about 25%, and compound repeats about 10%. Weber theorizes that because perfect repeats contain the longest uninterrupted repeat blocks, they appear to provide the most useful information. Weber also teaches that repeats composed of 12 or more uninterrupted units are consistently more polymorphic than are shorter repeat stretches. Because imperfect repeats generally contain shorter repeat stretches, they appear to be less useful as indicators of polymorphism. Compound repeats in general have not been well characterized, and their potential informativeness has not been clearly established.

Others have used the polymorphisms detectable within perfect, imperfect and compound SSR loci to build genetic linkage maps. Buchanan et al. (*Mammalian Genome*, 4, 258, (1993)) teach that there is little difference in the utility of the different SSR types in the ovine genome with respect to their absolute polymorphism levels; the perfect, imperfect and compound repeats although likely present in the genome at differing frequencies (perfect and imperfect simple SSR's are more frequent than compound) were found to have similar average Polymorphism Information Content (PIC) values as defined by Botstein et al. (*Am. J. Hum. Genet.*, 32, 314, (1980)). In a study of (GT)_n SSR in the Atlantic salmon genome, Slettan et al. (*Animal Genetics*, 24, 195, (1993)) found both perfect and imperfect simple SSR but no compound repeats. In an examination of the equine genome, Ellegren et al. (*Animal Genetics*, 23, 133, (1993)) identified the highest levels of polymorphism involving (TG)_n and (TC)_n repeats among horse genotypes using primers designed to amplify perfect or imperfect simple repeats; although two of eight cloned (GT)_n repeats were identified to be compound in structure (one perfect, one imperfect), neither was characterized further. Condit & Hubbell (*Genome* 34, 66 (1991)), in characterizing large-insert clones carrying (AC)_n and (AG)_n repeats from tropical trees and maize, found that 10-20% of inserts carrying one type of repeat also carried the other, and that many (AC)_n sites also had other two-base repeats adjacent or nearby. Finally, Browne et al. (*Nucl. Acids Res.*, 20, 141, (1991)), in an attempt to characterize (CA)_n SSR sequences in the human genome by DNA sequencing with degenerate (CA)_n primers, disclose that 88% of their (CA)_n repeats carried AT base pairs at one or both ends of the CA repeat.

To date, the record in the literature would indicate that although it varies with each type of genome, the incidence of compound SSRs in a genome is lower than that of either perfect or imperfect simple SSR sequences. Nevertheless, the information content (PIC value) of compound SSR sequences has been shown to be generally high. In addition, the literature would indicate that the detection of genetic polymorphisms by way of specifically isolating compound SSR loci generally would have marginal success; the use of probes or primers designed to recognize and thus specifically isolate individual compound SSR loci would be less efficient for generating large numbers of new SSR markers as compared to the isolation of the more numerous simple SSR sequences. Applicants have, however, unexpectedly discovered that compound SSR's, particularly those containing (AT)_n repeats are highly polymorphic in eukaryotic genomes, and that oligonucleotides designed to anneal specifically to a specific type of SSR, termed herein as a perfect in-phase compound SSR, are particularly useful in the generation and detection of polymorphisms between eukaryotic genomes. A "perfect" compound SSR is one in which two different repeating sequences, each of which could be composed of di-, tri-, tetra-, or penta-nucleotide units, are located very near each other, with no more than 3 intervening bases between the two repeat blocks. One category of perfect compound repeat is one in which the two constituent repeats are immediately adjacent to one another, with no intervening bases. Further, a perfect compound SSR can be classified to be "in-phase" if both of the component simple repeats share a common nucleotide whose spacing is conserved across the repeat junction and over the length of the two repeat blocks. For example, the in-phase perfect compound SSR, (AT)_n(AG)_n, maintains the adenosine base "in-phase" across both components of this perfect compound structure.

Applicants have discovered that in-phase perfect compound SSR sequences such as (dC-dA)_n(dT-dA)_n are abundant in both animal and plant genomes. Although the frequency of occurrence of each type of perfect compound SSR sequence varies within, as well as between, species, those sequences that are in-phase are of sufficiently high frequency in all eukaryotic genomes examined, and appear to be both well dispersed and highly polymorphic. Based upon their observation that the junction spanned by such directly adjacent, in-phase perfect repeats is absolutely predictable, Applicants have developed methodology which utilizes synthetic oligonucleotides containing in-phase compound sequences as self-anchoring primers in new variations of polymerase chain reaction-based multiplexed genome assays, including inter-repeat amplification and amplified fragment length polymorphism assays. Applicants have found that the 5' end of the compound SSR primer serves as an extremely efficient anchor base for primer extension that occurs from the 3'-end repeat. This primer extension initiates from inside the compound SSR target sequence, such

that any length variation between different alleles at a target SSR locus is detectable as a corresponding length variation in the resulting amplification products. Because such use of perfect compound SSRs as amplification primers generates multiple products wherein a high proportion are polymorphic (as high as 80%), Applicants believe that the method of their invention greatly facilitates the simultaneous identification of multiple genomic polymorphisms, both codominant and dominant. Thus, the present method offers great advantage in identifying polymorphic markers linked to genetic traits of interest, and also offers an efficient and convenient generic technique for genome fingerprinting and whole-genome comparisons.

SUMMARY OF THE INVENTION

This invention provides an improved method of detecting polymorphisms between two individual nucleic acid samples comprising amplifying segments of nucleic acid from each sample using primer-directed amplification and comparing the amplified segments to detect differences, the improvement comprising wherein at least one of the primers used in said amplification consists of a perfect compound simple sequence repeat. In a preferred embodiment, the compound primer is in-phase.

In a most preferred embodiment the present invention provides a method for the detection of genetic polymorphisms using a combination of in-phase perfect compound SSR primers and synthetic adaptor-directed primers for PCR amplification from restriction enzyme digested genomic DNA templates to which fixed-sequence adaptors have been ligated.

The present methods are particularly useful in the areas of clinical genetic diagnostics, forensic medicine (where it is important to detect small polymorphic changes in nucleic acid composition), as well as in the areas of animal and plant breeding and gene mapping. As specific applications, these methods have great utility for genome fingerprinting, polymorphic marker identification (i.e., "marking" a phenotypic trait), and germplasm comparisons.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates a schematic representation of SSR-to-adaptor amplification. Panel a depicts the restriction enzyme double digestion, adaptor ligation, and biotin-streptavidin selection process for generating the DNA template mixtures, as well as the selective nature of a specific SSR primer for exponential SSR-to-adaptor amplification from a complex template DNA mixture. Panel b depicts the selective nature of the adaptor-directed primer, in this case containing at its 3' end one nondegenerate selective nucleotide, for discriminating template fragments that otherwise would be amplified by a common SSR primer. Diagonally hatched boxes indicate the biotinylated adaptor corresponding to a restriction enzyme with a 6-bp recognition site, and dark boxes depict the nonbiotinylated adaptor corresponding to an enzyme with a 4-bp recognition site. The biotin moiety is indicated by a solid circle. The vertically striped box indicates either a simple SSR or a compound SSR of a particular type that matches the SSR primer used for the PCR amplification. Arrows depict PCR primers, with the arrowhead showing the direction of primer extension; solid/dark arrows indicate adaptor-directed PCR primers, and vertically striped-hatched arrows indicate primers corresponding to the SSR sequence depicted on the template fragments. Only the SSR-directed primer is tagged with either a fluorescent or radiolabel, as indicated by *. Panel C depicts a perfect, in-phase compound SSR as a double-stranded locus in the genome $((AT)_x(AG)_y)$, here where $x=11.5$ and $y=10$, which can serve as a target site for two classes of primer, each representing one strand of the double-stranded target locus. Individual primers within each class can differ by the relative length of each constituent repeat. The two classes of primer initiate primer extension in opposite directions (small arrows). In every case, the primer anneals to a fixed site at the target, and primer extension initiates inside the SSR region. In each case, any length variation in the 3'-most repeat between genomes could be detected as a codominant polymorphism using SSR-to-adaptor amplification.

Figure 2 illustrates an autoradiograph of a denaturing polyacrylamide gel that compares the co-amplified products from SSR-to-adaptor reactions on Taq I + Pst I digested, adaptor modified, biotin-selected template DNAs prepared from four different *Glycine max* or *Glycine soja* genotypes (N, *max* N85-2176; No, *max* NOIR-1; W, *max* wolverine; S, *soja* PI 81762). ^{32}P -labeled simple SSR primers containing 3 bp degenerate anchors at their 5' ends [HBH(AG) $_{8.5}$, DBD(AC) $_{7.5}$, HVH(TG) $_{7.5}$] are paired with unlabeled Taq I adaptor-directed primers containing either zero (TaqAd.F) or one (Taq.pr6, Taq.pr8) selective nucleotide at their 3' ends. Cold start amplifications employed either a constant temperature (58°C) or touchdown (59°C final temperature) thermocycle profile (left and right panels, respectively). An arrow indicates a likely codominant polymorphism.

Figures 3a and 3b illustrate autoradiographs of denaturing polyacrylamide gels that compare the co-amplified products from SSR-to-adaptor reactions on Taq I + Pst I digested, adaptor-modified, biotin-selected templated DNAs prepared from 15 different soybean genotypes. ^{32}P -labeled primers corresponding to perfect compound SSRs, (TC) 4.5(TG)4.5, (CT) $_{7.5}$ (AT) $_{3.5}$ and (CA) $_{7.5}$ (TA) $_{2.5}$ [panel a] or (TG) $_{4.5}$ (AG) $_{4.5}$ and (TC) $_{4.5}$ (AC) $_{4.5}$ [panel b], each are paired with unlabeled Taq I-adaptor primers containing either zero (TaqAd.F) or one (Taq.pr8) 3'-selective nucleotide, under

cold start amplification conditions utilizing a touchdown (56°C final) thermocycle profile. In each set, lane 1, *Glycine max* wolverine; lane 2, *G. max* NOIR-1; lane 3, *G. max* N85-21761; lane 4, *G. max* Harrow; lane 5, *G. max* CNS; lane 6, *G. max* Manchu; lane 7, *G. max* Mandarin; lane 8, *G. max* Mukden; lane 9, *G. max* Richland; lane 10, *G. max* Roanoke; lane 11, *G. max* Tokyo; lane 12, *G. max* PI 54.610; lane 13, *G. max* Bonus; lane 14, *G. soja* PI 81762; lane 15, *G. soja* PI 440.913. The size distribution of products is similar to that in Figure 2. Lane 9 of the (TG)_{4.5}(AG)_{4.5}+Taq.Pr8 set is a misloading of an incorrect (non-soybean) sample.

Figure 4 illustrates an autoradiograph of a denaturing polyacrylamide gel that compares the co-amplified products from SSR-to-adaptor amplifications performed on five different soybean genotypes (lane 1, *G. max* wolverine; 2, *G. max* NOIR-1; 3, *G. max* N85-2176; 4, *G. max* Bonus; 5, *G. soja* PI 81762) prepared by digestion with Taq I combined with either Hind III or Pst I (H+T or P+T respectively). Cold start amplifications using a 56°C touchdown thermocycle profile utilized the ³³P-labeled SSR primer, (CA)_{7.5}(TA)_{2.5}, in combination with the indicated unlabeled Taq I adaptor-directed primer, Taq.Ad.F or Taq.Pr8 (zero or one 3'-selective nucleotide, respectively). X indicates a misloaded (incorrect) lane.

Figure 5a illustrates an autoradiograph of a denaturing polyacrylamide gel demonstrating the segregation of polymorphic co-amplification products of 66 F2 progeny from a cross between *G. soja* PI 81762 and *G. max* Bonus. A ³³P-labeled primer corresponding to the perfect compound SSR, (CA)_{7.5}(TA)_{2.5}, was paired with the 3'-selective nucleotide Taq.pr6 adaptor-directed primers. Blank lanes are the result of "missing data". The scored polymorphic bands that segregate in this population are indicated. B, bonus parent; S, *soja* PI81762 parent. Figure 5b illustrates the map positions of 6 of the polymorphic segregating amplification products contained in panel a, as determined by MAPMAKER analysis of the products' respective segregation scores (contained in Table VI).

Figure 6a illustrates an autoradiograph of a denaturing polyacrylamide gel comparing amplifications using the perfect compound SSR primer (CA)_{7.5}(TA)_{2.5}, paired with either the Taq.AdF or the more selective Taq.pr8 adaptor-directed primer, on template DNAs derived from either 5 (wolverine, NOIR-1, N85-2176, Bonus, PI 81762) or 15 soybean cultivars same ordering of genotypes as in Figures 3a and 4, respectively, and 6 mammalian individuals (one rat, four human, one mouse BALB/C). All biotin-selected templates were prepared using Taq I combined with either Hind III or Pst I. Figure 6b illustrates a similar comparison of the co-amplification products using (CT)_{7.5}(AT)₂ and (GA)_{7.5}(TA)₂ perfect compound SSR primers, from Pst I + Taq I prepared template DNAs. The size distribution of products is similar to that in panel a. Figure 6c illustrates the SSR-to-adaptor amplification products generated from Taq I + Hind III prepared templates of *Zea mays* (corn) and salmon DNA templates, in comparison to those from soybean Bonus and PI81762, using (CA)_{7.5}(TA)_{2.5} paired with Taq.pr6 primer. Lane 1, *Z. mays* B73; lanes 2 and 3, individual salmon sources; lane 4, *Z. mays* CM27; lane 5, *Z. mays* T232; lane 6, *Z. mays* DE811ASR; lane 7, *Z. mays* LH132; lanes 8-10, two F2 individuals from a *G. max* Bonus x *G. soja* PI 81762 cross. The distribution of product sizes is similar to that shown in Figure 6a.

Figure 7 illustrates an autoradiograph of a denaturing polyacrylamide gel comparing the co-amplification products from Taq I + Pst I prepared soybean templates (S, *Soja* PI81762; W, *Wolverine*; B, bonus) using (CA)_{7.5}(TA)_{2.5} paired individually with Taq I adaptor-directed primers carrying either zero (Taq.AdF) or one specific 3'-selective nucleotide (Taq.pr5, .pr6, .pr7, .pr8).

Figure 8 illustrates an autoradiograph of a denaturing polyacrylamide gel comparing SSR-to-adaptor amplification from wolverine and PI 81762 soybean cultivars using SSR primers representing the complementary strands of a perfect compound SSR double stranded sequence [(AT)_x(GT)_y·(CA)_x(TA)_y] paired separately with three different Taq I adaptor-directed primers, Taq.AdF, Taq.pr6 and Taq.pr8. Each strand of the double stranded compound SSR sequence is represented by three primers that differ by the relative lengths of each of the two constituent repeats within the primer. The (AT)_{8.5}(GT)_{2.5} and (AT)_{6.5}(GT)_{4.5} primers were completely inefficient (no amplification products generated (data not shown) and (AT)_{3.5}(GT)_{6.5} was moderately successful (shown), in comparison to the three, more efficient (CA)_x(TA)_y primer types.

Figure 9 illustrates autoradiographs of denaturing polyacrylamide gels that compare the co-amplification products from soybean cultivars, wolverine and PI 81762, using cold start (left) and hot start (right) methods for the initiation of thermocycling. For both methods, the perfect compound SSR primer, (CA)_{7.5}(TA)_{2.5}, is paired with each of the three Taq I adaptor directed primers indicated.

Figure 10a illustrates an autoradiograph of a denaturing polyacrylamide gel comparing the co-amplification products from soybean (B, Bonus; S, *soja*; PI81762) and corn (b, B73; c, CM37) cultivars amplified using DBD(AC)_{6.5} in combination with no 2nd adaptor primer or with Taq.pr8. These templates are in the form of either intact, undigested DNA or Taq I + Pst I digested DNA (P+T), as indicated.

Figure 10b illustrates autoradiographs of denaturing polyacrylamide gels comparing the amplification products obtained using ³³P-labeled 5'-anchored simple SSR primers [DBD(AC)_{7.5} and HBH(AG)_{8.5}] or perfect compound SSR primers [(AT)_{3.5}(AG)_{7.5} and (AT)_{3.5}(GT)_{6.5}] in single-primer amplifications (no adaptor primer used) from both undigested and Taq I + Pst I digested, biotin-selected template DNAs from soybean wolverine (W) and PI81762 (S) cultivars. Cold start amplifications used either constant temperature (58°C) or 56°C touchdown annealing profiles.

Figure 11 is a schematic representation of the cloning and sequencing of a chosen SSR-to-adaptor amplification product, and its conversion into a defined, single-locus marker. A genomic restriction fragment carrying the targeted SSR repeat is bordered at both ends by restriction site-specific adaptors, Ad_A and Ad_B. This fragment serves as the template for PCR amplification using an SSR-directed primer and a primer corresponding to one of the adaptors. This amplification product is purified and sequenced, and a locus-specific flanking primer (lsfp-1) is designed. This lsfp-1 primer then is paired with a primer corresponding to the other adaptor, for PCR amplification using the adaptor-modified restriction fragment mixture as template. The specific product obtained is then isolated and sequenced, and a second primer (lsfp-2) corresponding to the unique flanking sequence on the other side of the SSR is designed. The lsfp-1 + lsfp-2 primer pair uniquely defines this SSR locus, and can be used to amplify directly from genomic DNA to visualize SSR length polymorphism at this locus. Applicants have provided 89 sequence listings in conformity with 37 C.F.R. 1.821-1.825 and Appendices A and B ("Requirements for Application Disclosures Containing Nucleotides and/or Amino Acid Sequences").

DETAILED DESCRIPTION OF THE INVENTION

As used herein the following terms may be used for interpretation of the claims and specification.

The term "Simple sequence repeat (SSR)" or "microsatellite repeat (MS)" or "short tandem repeat" or "dinucleotide repeat" or "trinucleotide repeat" or "tetranucleotide repeat" or "microsatellite" or "simple sequence length polymorphisms" (SSLP) all refer to stretches of DNA consisting of tandemly repeating di-, tri-, tetra-, or penta-nucleotide units. An SSR region can be as short as two repeating units, but more frequently is in excess of 8-10 repeating units. Simple sequence repeats are common in virtually all eukaryotic genomes studied and have been identified as useful tools for the study of genetic polymorphisms.

Classification of SSR loci or SSR sequences as used herein is based upon (but not identical to) the definitions suggested by Weber (*Genomics* 7, 524 (1990)) for the categorization of human (CA)_n dinucleotide repeats.

The term "simple SSR" will refer to a region comprised of at least three or more of the same tandemly repeated di-, tri- or tetranucleotide sequence, which is not adjacent in the genome to any other different simple SSR. "Not adjacent" means not closer than four nucleotides away on either side.

The term "compound SSR" refers to a region consisting of two or more different simple SSR sequences which are adjacent. "Adjacent" means that differing simple SSR's are separated from one another by three or fewer consecutive nonrepeat nucleotides.

The term "perfect SSR" refers to a simple SSR wherein every simple repeating unit within the SSR is intact and uninterrupted by nonrepeat nucleotides.

The term "imperfect SSR" refers to a simple SSR wherein one or more of the constituent repeat units is interrupted at least once within the SSR by three or four consecutive nonrepeat nucleotides.

The term "perfect compound SSR" refers to a compound SSR wherein the two constituent repeating SSR regions are intact and uninterrupted by non repeat bases, (i.e., they are perfect SSR's), and the two perfect SSR regions are located very near each other, with no more than 3 intervening bases between the two repeat blocks, for example directly adjacent to one another, having no intervening nucleotides.

The term "in-phase" refers to a potential feature of a perfect compound SSR wherein both constituent SSR regions share a common nucleotide that retains constant spacing spanning the junction of the two SSR regions.

The term "out of phase" refers to a potential feature of a perfect compound SSR wherein a nucleotide is common to the two or more constituent repeating regions, but it does not retain constant spacing or periodicity across the junction of the compound structure.

The term "polymorphism" refers to a difference in DNA sequence between or among different genomes or individuals. Such differences can be detected when they occur within known or tagged genomic regions. A "dominant polymorphism" is a DNA difference that is detectable only as the presence or absence of a specific DNA sequence at a single locus. Methods to detect dominant polymorphisms are able to detect only one allele of the locus at a time, and genomes homozygous versus heterozygous for the detectable allele are indistinguished. A "codominant polymorphism" is a DNA difference at a locus between genomes whereby multiple alleles at the locus each can be distinguishable even when in heterozygous combinations. Typically identifiable as mobility variants on electrophoretic gels, codominant polymorphisms can produce additive, nonparental genotypes when present in heterozygous form. A dominant polymorphism is most useful as a marker when it is in coupling with the trait it marks, whereas a codominant polymorphism is equally useful when in coupling or in repulsion to a trait.

The term "touchdown amplification" or "touchdown PCR" will refer to a specific thermocycling profile for the polymerase chain reaction whereby the annealing temperature begins artificially high (or low) for the first few cycles, then is incrementally lowered (or raised) for a specified number of successive cycles until a final, desirable annealing temperature is reached. The remaining cycles of the multiple cycle profile are then performed at this final, touchdown annealing temperature. Thermocycling using this strategy serves to reduce or circumvent spurious, nonspecific priming

during the initial stages of gene amplification, and imbalance between correct and spurious annealing is automatically minimized.

The terms "hot start" and "cold start" will refer to a general choice of methodologies for initiating thermocycling for a PCR amplification reaction. In a cold start amplification, all the reaction components are assembled simultaneously at room temperature, prior to the first denaturation step. This approach allows for the possibility of spurious priming and nonspecific amplification products resulting from primer annealing and primer extension at undesirably low temperatures. In contrast, a hot start approach employs the deliberate omission of at least one key component from the otherwise complete amplification reaction, thus preventing either primer annealing (if primer is omitted) or extension (if either polymerase or nucleotides are omitted). After carrying out an initial high-temperature denaturation, the excluded component is added, and thermocycling proceeds. A hot start amplification thus serves to reduce or eliminate the production of nonspecific products that result from spurious primer extension at nonstringent temperatures.

"Nucleic acid" refers to a molecule which can be single stranded or double stranded comprised of monomers (nucleotides) containing a sugar, phosphate and either a purine or pyrimidine. In bacteria, lower eukaryotes, and in higher animals and plants, "deoxyribonucleic acid" (DNA) refers to the genetic material while "ribonucleic acid" (RNA) is involved in the translation of the information from DNA into proteins.

The terms "genomic DNA" or "target DNA" or "target nucleic acid" will be used interchangeably and refer to nucleic acid fragments targeted for amplification or replication and subsequent analysis by the instant method for the presence of SSR regions. Sources of genomic DNA will typically be isolated from eukaryotic organisms. Genomic DNA is amplified via standard replication procedures using suitable primers to produce detectable primer extension products.

The term "restriction endonuclease" or "restriction enzyme" is an enzyme that recognizes a specific palindromic-base sequence (target site) in a double-stranded DNA molecule, and catalyzes the cleavage of both strands of the DNA molecule at a particular base in every target site.

The term "restriction fragments" refers to the DNA molecules produced by digestion with a restriction endonuclease. Any given genome may be digested by a particular restriction endonuclease into a discrete set of restriction fragments. The DNA fragments that result from restriction endonuclease cleavage may be separated by gel electrophoresis and detected, for example, by either fluorescence or autoradiography.

The term "restriction fragment length polymorphism (RFLP)" refers to differences in the genomic DNA of two closely related organisms which are detected based upon differences in the pattern of restriction fragments generated by a restriction endonuclease digestion of genomic DNA of the organisms. For example, a genome which contains a polymorphism in the target site for a restriction endonuclease will not be cleaved at that point by the restriction endonuclease. Or, a nucleotide sequence variation may introduce a novel target site where none exists in the other organism, causing the DNA to be cut by the restriction enzyme at that point. Additionally, insertions or deletions of nucleotides occurring between two target sites for a restriction endonuclease in the genome of one organism will modify the distance between those target sites. Thus, digestion of the two organism's DNA will produce restriction fragments having different lengths and will generate a different pattern upon gel electrophoresis.

The term "ligation" refers to the enzymatic reaction catalyzed by the enzyme T4 DNA ligase by which two double-stranded DNA molecules are covalently joined together in their sugar-phosphate backbones via phosphodiester bonds. Ligation can occur between two DNA molecules that each are bounded by blunt (nonstaggered) ends, but also can occur if the two DNA molecules contain single-stranded overhanging ends that are complementary in sequence. In general, both DNA strands of the two double helices are covalently joined together such that at each junction the free 5' end of one of the DNA molecules carries a 5'-phosphate group. It is also possible to prevent the ligation of one of the two strands, through chemical or enzymatic modification (for example, removal of the 5' phosphate) of one of the ends, in which case the covalent joining would occur in only one of the two DNA strands.

The term "adaptor" will specifically refer herein to short, largely double stranded DNA molecules comprised of a limited number of base pairs, e.g., 10 to 30 bp. Adaptors are comprised of two synthetic single-stranded oligonucleotides having nucleotide sequences that are not intentionally represented by repetitive sequences in the genome of interest, and also are in part complementary to each other. Under appropriate annealing conditions, the two complementary synthetic oligonucleotides will form a partially double-stranded structure in solution. At least one of the ends of the adaptor molecule is designed so that it is complementary to and can be specifically ligated to the digested end of a restriction fragment.

The term "polymerase chain reaction" or "PCR" refers to the enzymatic reaction in which copies of DNA fragments are synthesized from a substrate DNA in vitro (U.S. Pat. Nos. 4,683,202 and 4,683,195). The reaction involves the use of one or more oligonucleotide primers, each of which is complementary to nucleotide sequences flanking a target segment in the substrate DNA. A thermostable DNA polymerase catalyzes the incorporation of nucleotides into the newly synthesized DNA molecules which serve as templates for continuing rounds of amplification.

The term "DNA amplification" or "nucleic acid amplification" or "nucleic acid replication" or "primer extension" refers to any method known in the art that results in the linear or exponential replication of nucleic acid molecules that are copies of a substrate DNA molecule.

The term "primer" refers to a DNA segment that serves as the initiation point or site for the replication of DNA strands. Primers generally will be single-stranded and will be complementary to at least one strand of the target or substrate nucleic acid and will serve to direct nucleotide polymerization or primer extension using the targeted sequence as a template. Primers may be used in combination with another primer to "flank" the target sequence in PCR, thus forming a "primer set" or "primer pair". In general, primers are 14 to 40 nucleotides long and preferably are designed so as not to form secondary structure or hairpin configurations. Specific requirements for primer size, base sequence, complementarity and target interaction are discussed in the primer section of the detailed description of the invention. The term "primer", as such, is used generally herein by Applicants to encompass any synthetic or naturally occurring oligonucleotide that can hydrogen-bond specifically to a region of a substrate DNA molecule and functions to initiate the nucleic acid replication or primer extension process; such processes may include, for example, PCR, or other enzymatic reactions that employ single rather than multiple oligonucleotide initiators.

The term "anchor" or anchor region" or "anchor portion" refers to a 3-20 nucleotide region of a primer designed to hybridize with a DNA sequence which is immediately adjacent to a specified sequence SSR. The anchor region of a primer may occur at either the extreme 5' or 3' end, and serves to affix the primer onto the target DNA at an adjacent position relative to a specified SSR. This anchoring results in primer extension occurring from a fixed nucleotide at each target site. The anchor sequence can be a nondegenerate sequence of either deliberate or arbitrary design, or it can be a fully or partially degenerate sequence. The latter would be capable of annealing to the genomic DNA sequences flanking a wide range of SSR sites in a genome. Optionally, the anchor portion of the primer may be 5' end-labeled with a reporter molecule, typically a radioisotope, a fluorescent moiety or a reactive ligand.

The use of the term "arbitrary" when speaking of an individual nucleotide at each position in a DNA sequence refers to selection based on or determined by unbiased means or seemingly by chance rather than by necessity or by adherence to a predetermined sequence.

The term "non-degenerate" refers to the occurrence of a single, specified nucleotide type at a particular position or at multiple specified positions in the linear ordering of nucleotides in a DNA polymer, usually an oligonucleotide or a polynucleotide. Any nondegenerate nucleotide position can carry an intended base (either A, G, C or T) that is known for example to correspond to a given template site, or it can carry an arbitrarily chosen base, which will correspond to a target site that is not known *a priori*. A "non-degenerate oligonucleotide" means that every nucleotide position within the DNA molecule is non-degenerate. The term "degenerate" refers to the occurrence of more than one specified nucleotide type at a particular position in an oligonucleotide or polynucleotide. A specific oligonucleotide can be made up of some positions that are degenerate and some positions that are fully or partially degenerate. "Fully degenerate" indicates the presence of an equal mixture of the four possible nucleotide bases (A, G, C or T) at a particular nucleotide position; partially degenerate indicates the presence of only two or three of the four possible bases at a particular position. A "degenerate oligonucleotide" is one in which at least one position within it carries full or partial degeneracy; such an oligo- or polynucleotide is a mixture of specific, nondegenerate DNA molecules, each of which represents a single permutation of the nucleotide sequences possible by virtue of the degenerate base(s) specified in the linear nucleotide sequence. An oligonucleotide with two fully degenerate positions, for example, would be a mixture of $(4)^2=16$ different nondegenerate molecules; an oligonucleotide with four fully degenerate and two partially degenerate (three bases) positions would comprise a mixture of $(4)^3 \times (3)^2=576$ different non-degenerate molecules. Standard degeneracy codes used herein are:

N or X	A, G, C or T
B	G, C or T [anything except A]
D	A, G or T [anything except C]
H	A, C or T [anything except G]
V	A, C or G [anything except T]

The term "reporter" or "reporter molecule" refers to any moiety capable of being detected via enzymatic means, immunological means or energy emission; including, but not limited to, fluorescent molecules, radioactive tags, light emitting moieties or immunoreactive or affinity reactive ligands.

The term "binding pair" includes any of the class of specific inter-molecular or recognition immune-type binding pairs, such as antigen/antibody or hapten/anti-hapten systems; and also any of the class of nonimmune-type binding pairs, such as biotin/avidin; biotin/streptavidin; folic acid/folate binding protein; complementary nucleic acid segments; protein A or G/immunoglobulins; and binding pairs which form covalent bonds, such as sulfhydryl reactive groups including maleimides and haloacetyl derivatives, and amine reactive groups such as isothiocyanates, succinimidyl esters and sulfonyl halides.

The present invention describes the design and use of self-anchoring primers for the detection of SSR genetic markers. The general polymorphism detection method using these primers is termed selective amplification of micro-

satellite polymorphic loci (SAMPL). The method of primer design is based on the observation that many compound SSR sequences are composed of dinucleotide repeats wherein a single type of nucleotide is shared by both of the directly adjacent constituent repeats and is maintained "in-phase" across the repeat junction and throughout the length of the repeat. The present invention combines many of the advantages inherent to conventional, single-locus SSR markers (i.e., high levels of polymorphs and high codominance potential), with the added benefits and convenience offered by multiplexed genome assays. As with conventional single-locus SSR markers, the use of perfect compound SSR sequences as self-anchored PCR primers enables the identification of dominant and codominant polymorphisms between genomes. However, unlike conventional SSR analysis, this method requires no prior knowledge of the unique sequences flanking individual SSR loci. Thus, no labor-intensive SSR marker discovery or locus identification is necessary to use such compound SSR sequences as primers as we describe. Also, as with conventional SSR markers, the in-phase subset of compound SSR sequences appears to be highly abundant and well dispersed in both plant and animal genomes, as well as to be highly polymorphic between individual genomes. In contrast, the "out-of-phase" subset of perfect compound SSR sequences are represented in these same genomes at much lower relative frequencies. Therefore, the likelihood that the highly abundant, in-phase perfect compound SSR sequences can identify new polymorphisms closely linked to loci of interest is extremely high. Additionally, PCR primers representing these compound SSR sequences are self-anchoring, such that the 5'-most repeat serves as the anchor for primer extension by the 3'-most of the two repeats, thus obviating the need to incorporate into these primers any additional degenerate or fixed sequences as 5' or 3' flanking anchors. Therefore, conceivably every genomic locus harboring the same compound SSR sequence would be expected to serve as a target site for simultaneous amplification by the single compound SSR primer matching these target sites. Finally, and most preferred, the use of these primers can be incorporated into several different genome assays to increase their versatility and informativity. For example, the use of these perfect, in-phase compound SSR primers in modifications of the amplified fragment length polymorphism assay (AFLP; Zabeau, EP 534,858) leads to an increase in the proportion of amplification products that are polymorphic and codominant between even highly related genomes as compared to conventional AFLP methods. The versatility of these compound SSR primers, in combination with the ability to fine-tune both the numbers and types of amplification products achievable with the AFLP assay, offers a unique combination of benefits for the multiplexed analysis of complex plant and animal genomes.

Applicants' modified AFLP invention is illustrated in Figure 1. Genomic DNA (I) carrying perfect in-phase compound SSR sequences at some frequency of occurrence is digested with a single restriction enzyme, or with a combination of two or more restriction enzymes. This Figure demonstrates the use of a double enzyme combination, one enzyme having a hexanucleotide recognition site (hatched boxes) and the other a tetranucleotide site (dark boxes). It is also within the scope of the invention to choose other combinations of multiple restriction enzymes, including but not limited to the following combinations of restriction enzymes with the specified types (lengths) of recognition site:

35

40

45

<i>Additional Enzyme combinations</i>	
two enzymes	three enzymes
4 + 4	4 + 4 + 4
5 + 5	5 + 4 + 4
6 + 6	6 + 4 + 4
4 + 5	5 + 4 + 5
5 + 6	6 + 4 + 5
4 + 8	6 + 4 + 6
5 + 8	5 + 5 + 5
	5 + 5 + 6
	6 + 6 + 6

In all cases, these multiple enzyme digestions produce a mixture of restriction fragments with all combinations of the corresponding blunt or single stranded overhanging ends. Additionally it is within the scope of the invention for a single restriction enzyme to be used to produce fragments all sharing the same blunt or single-stranded overhanging ends. In general, any enzyme with a 4-, 5-, 6- or 8-bp recognition site is suitable, providing the enzyme's activity is not affected or inhibited by DNA methylation or other, nonmendelian modes of DNA modification within the enzyme site.

Next, double stranded adaptors A and B are constructed wherein Adaptor A anneals specifically to the single stranded overhang produced by the hexanucleotide-site enzyme, and Adaptor B to the overhang left by the tetranucleotide-site enzyme. Adaptors A and B are simultaneously ligated to the appropriate ends of all the restriction fragments (II) in the digested DNA mixture using standard methods (also described in Table III). In an alternate embodiment, either Adaptor A or Adaptor B may be conjugated with a member of a binding pair such as biotin (as part of a biotin-

streptavidin pair), allowing capture and isolation of a smaller subset of the genomic restriction fragments (III). Either of the adaptors can be so modified; the proportion of genomic fragments then selected for is affected by the genomic frequency of the specific restriction site recognized by the modified adaptor. For any given restriction enzyme combination, this enrichment method categorically allows only a small fraction of the total number of restriction fragments from a genome the opportunity to serve as template for the subsequent PCR amplification; however, this reduced complexity is necessary for ensuring a manageable number of co-amplified products in the next step. The entire genome can be examined through the use of multiple combinations of restriction enzymes for the generation of different sets of enriched genomic fragments. Figure 1a illustrates a biotinylated hexanucleotide site adaptor.

Alternatively, the complexity of the genomic fragment mixture can be selectively reduced by performing a pre-amplification step prior to the final amplification, using a pair of unlabeled adaptor-directed primers. This primer pair comprises two different primers, one corresponding to one adaptor sequence and the second primer to the other adaptor sequence. Each primer carries one selective nucleotide at its 3'-end. Since the 3'-most position of each of these +1 adaptor primers can be occupied by any one of the four DNA nucleotides, each adaptor can be represented by 4 different primers. Furthermore, $4 \times 4 = 16$ different combinations of these +1 primers can be used against any genomic fragment mixture to generate 16 different, nonoverlapping fragment subsets from each genome. Thus, the pre-amplification enriches for a subset of the total mixture of genomic fragments, and different enriched subsets can be generated from a single restriction fragment mixture by varying the specific primers used for the pre-amplification.

No matter how the fragment enrichment is performed, a pair of PCR primers next are synthesized according to standard protocols. One primer will be an adaptor-directed primer, designed to anneal to adaptor B specifically when the biotin-mediated fragment enrichment method is employed. For enrichment via pre-amplification, this primer can correspond to either of the two adaptors. In either case, this adaptor primer carries 1-4 randomly selected nucleotides at their 3'-ends.

The other primer will be a SSR-directed primer, designed to anneal specifically to a particular SSR sequence represented on a subset of the genomic fragments. In a preferred embodiment, the SSR-directed primer is 5'-end labeled, typically with a radionucleotide such as ^{32}P or ^{33}P or with a fluorescent moiety (*). It is further especially preferred if the SSR-directed primer is of the "perfect compound" type wherein the primer straddles the compound SSR, preferably with one nucleotide remaining "in-phase" across the length of the primer. Exponential amplification of a subset of the adaptor modified restriction fragments in the presence of the adaptor-directed primer and the 5'-labeled SSR primer generates labeled primer extension products (IV) from every input genomic fragment that carries the SSR sequence and is bordered at the opposing end by the designated adaptor sequence.

This method generates multiple co-amplification products, a high proportion of which are expected to be polymorphic between genomes. Those genomic fragments lacking either the designated SSR sequence or the appropriate adaptor end, or both, will not be exponentially amplified, and therefore will not be detected (V).

GENERAL METHODS:

Primer Design:

All oligonucleotides primers are synthesized using solid-phase phosphoramidite chemistry such as that described by Operon Technologies, Alameda, CA. All primers are non-phosphorylated at their 5' ends and can be used in unpurified form providing efficient syntheses. However, oligonucleotides purified by column chromatography are preferred for optimal primer specificity. All oligonucleotide primer sequences are chosen such that the T_m in 50 mM salt (KCl or NaCl) is between 38° and 45°C (as determined using the algorithm employed by Oligo v4.03 for the Macintosh, National Biosciences, Inc.).

Several types of primers are used within the context of Applicants' invention. The first is an adaptor-directed primer (as disclosed in Zabeau EP 534,858) which can vary from 15 to 25 nucleotides in length, and from 40% and 60% G+C. Starting at its 5' end, the primer spans the length of, and is complementary to, one strand of a double-stranded adaptor that is ligated to the restriction endonuclease-digested target DNA to be tested. The primer then covers all or part of the restriction site, and its 3' end can carry arbitrary, nondegenerate bases that anneal to and prime from nucleotides within the target DNA fragment adjacent to the adaptor. The sequence of such an adaptor-directed primer can vary, depending upon the specific adaptor used for the construction of the template and upon the number of arbitrary, selective nucleotides positioned at its 3'-end. Examples of DNA sequences and characteristics of oligonucleotides comprising both adaptor and adaptor-directed primers, each specific to the site generated by a particular restriction enzyme, are given in but not limited to Table I.

TABLE I
ADAPTOR AND PRIMER OLIGONUCLEOTIDES

Double Strand Adaptor	Oligonucleotide Name	Sequence(5'→3')/SEQ ID No.	Length	Selective 3' nts	T _m (50mM Salt)	%GC
Adaptor oligonucleotide components						
biotin-Pst I-Ad	B-Pst.Ad.F	B-CTCGTAGACTGCGTACATGCA /17	21	0	49.2°C	52.40
	Pst.Ad.R	3'-CATCTGACGCAATG-5' /18	14	0	25.7°C	50.00
biotin-Hnd III-Ad	B-Hnd.Ad.F	B-CTCGTAGACTGCGTACC /19	17	0	44.3°C	58.80
	Hnd.Ad.R	3'-CTGACGCAATGGTCA-5' /20	15	0	39.5°C	60.00
Taq I-Ad	Taq.Ad.F	GACGATGAGTCCTGAC /21	16	0	40.8°C	56.20
	Taq.Ad.R	3'-TACTCAGGACTGGC-5' /22	14	0	35.1°C	57.10
Sau-Ad	Sau.Ad.F	GGAATTCCTGGACTCAGT /23	17	0	39.5°C	47.00
	Sau.Ad.R	3'-CCTTAAGACCTGATCACTAG-5' /24	21	0	47.3°C	47.60
Mse-Ad	Mse.Ad.F	TGGCCTTTACAGCGTC /25	16	0	40.8°C	56.30
	Mse.Ad.R	3'-GAAATGTCGCACAT-5' /26	14	0	29.3°C	42.90
Adaptor-directed primers						
Hnd.Ad.F	B-CTCGTAGACTGCGTACC /27		17	0	44.3°C	58.80
Hnd.pr.1	CTGCGTACCAGCTTaca	/28	17	3 -aca	41.9°C	52.90
Hnd.pr.2	CTGCGTACCAGCTTacc	/29	17	3 -acc	44.3°C	58.80
Hnd.pr.3	CTGCGTACCAGCTTaac	/30	17	3 -aac	41.9°C	52.90
Hnd.pr.4	CTGCGTACCAGCTTgtc	/31	17	3 -gtc	44.3°C	58.80
Hnd.pr.5	CTGCGTACCAGCTTac	/32	16	2 -ac	40.8°C	56.20
Hnd.pr.6	CTGCGTACCAGCTTaa	/33	16	2 -aa	38.2°C	47.05

A second type of primer used within the present invention corresponds in its 3' portion to a simple sequence repeat, or microsatellite, where the structure of the microsatellite is simple (simple SSR) as defined above. The simple microsatellite region can be tandem repeats of mono-, di-, tri-, tetra- or pentanucleotides. The 5' position of the primer contains 3 to 5 fully or partially degenerate nucleotides, which serve to anchor the primer adjacent to a microsatellite in the targeted genome. This primer can vary in length from 10 to 60 nucleotides, with a [G+C] content typically from 16% to 80%. Length polymorphism within a microsatellite locus between genomes is expected to be detectable using these primers, since in nearly all cases, primer extension is expected to initiate from a fixed site relative to the SSR target region. Primers similar to these are described by Zietkiewicz, E., et al., *Genomics*, 20, 176, (1994).

Another type of SSR-directed primer, uniquely utilized within the present invention, is the perfect compound SSR primer which is comprised of two different perfect SSR's that are immediately adjacent to one another with no inter-

vening nucleotides either between the repeats or within each of the repeats. Perfect compound SSR primers are perfectly self-anchoring; that is, the simple SSR at the 5'-end of the primer serves as an efficient anchor for the adjacent 3' SSR from which primer extension proceeds (see Figure 1c). It is intended, therefore, that primer extension initiates from a single, fixed site within a compound SSR target region. Any length variation between genomes in the portion of the target SSR across which primer extension occurs should be visible as length variation in the resulting amplification products from those genomes. The relative lengths of each constituent SSR within the compound primer can vary. However, Applicants have found that the best primer anchoring and greatest specificity for the target template is produced when the length of the 5' anchor is equal to or greater than that of the 3' priming portion.

For dinucleotide repeats, Applicants theorize that, excluding CG and GC combinations (long stretches of which are thought to be rare in eukaryotic genomes), 90 different permutations of two adjacent dinucleotide sequences are possible. As estimated by the following equation,

$$[(4)(3)-2] \times [(4)(3)-2-1] = 90,$$

the first (5'-most) constituent repeat may carry of any of the four nucleotides (A, G, C or T) in its first position, followed by any of the three remaining nucleotides at its second position, then from this product should be subtracted the two GC and CG combinations. For the second (3'-most) dinucleotide, the same calculation holds, but with the additional subtraction of the one combination occupying the first constituent dinucleotide repeat position. All of these 90 permutations are listed in Table II. Only 80 are true compound repeat sequences, however; 10 of the permutations are actually imperfect simple repeats (e.g., $(CT)_n(TC)_n$, $(AG)_n(GA)_n$, etc.). Similar calculations can be performed to estimate the number and types of different tri-, tetra- and penta-nucleotide combinations possible by random nucleotide arrangements.

TABLE II

Occurrence in DNA Sequence Database

Permutations of Adjacent Dinucleotides:

COMPOUND, IN-PHASE Double Stranded Locus	Permutations 5' → 3'	Total, all databases	Primate + human EST	Rodent	Mammal	Other Vertebrate	Invertebrate	Bacteria	Virus	Phage	Fungi	Plant	Clinical Soybean
5'-(AC) _n (AT) _n -3' 3'-(TG) _n (TA) _n -5'	(AC)(AT)(CA)(TA) (AT)(GT)(TA)(TG)	79	28	27	5	2	8	0	0	0	5	4	22
5'-(AT) _n (AC) _n -3' 3'-(TA) _n (TG) _n -5'	(TA)(CA)(AT)(AC) (TG)(TA)(GT)(AT)	65	47	6	7	3	1	0	0	0	0	1	4
5'-(AG) _n (AT) _n -3' 3'-(TC) _n (TA) _n -5'	(AG)(AT)(GA)(TA) (AT)(CT)(TA)(TC)	0	0	0	0	0	0	0	0	0	0	0	1
5'-(AT) _n (AG) _n -3' 3'-(TA) _n (TC) _n -5'	(TA)(GA)(AT)(AG) (TC)(TA)(CT)(AT)	15	5	0	3	1	1	0	0	0	0	5	11
5'-(AC) _n (AG) _n -3' 3'-(TG) _n (TC) _n -5'	(AC)(AG)(CA)(GA) (CT)(GT)(TC)(TG)	83	40	41	2	0	0	0	0	0	0	0	3
5'-(AG) _n (AC) _n -3' 3'-(TC) _n (TG) _n -5'	(AG)(AC)(GA)(CA) (GT)(CT)(TG)(TC)	3	2	1	0	0	0	0	0	0	0	0	2
5'-(TG) _n (AG) _n -3' 3'-(AC) _n (TC) _n -5'	(TG)(AG)(GT)(GA) (CT)(CA)(CA)(CT)	221	98	43	71	2	4	0	2	0	0	1	5
5'-(AG) _n (TG) _n -3' 3'-(AC) _n (TC) _n -5'	(AG)(TG)(GA)(GT) (CA)(CT)(AC)(TC)	65	56	5	1	0	3	0	0	0	0	0	1
total occurrences # bp searched		531	276	123	89	8	17	0	2	0	5	11	49
# sequences searched		2.0E+08	3.3E+07	2.5E+07	6.7E+06	8.0E+06	2.2E+07	3.0E+07	2.2E+07	1.5E+06	3.1E+06	>>	>>
		2.0E+05	3.5E+04	2.2E+04	6.2E+03	7.4E+03	1.3E+04	1.7E+04	1.8E+04	9.9E+02	1.8E+04	>>	>>

5

10

15

20

25

30

35

40

45

50

55

COMPOUND, OUT-OF-PHASE Double Stranded Locus	Permutations 5' → 3'	Total, all databases	Note: x, y ≥ 6 for all dinucleotide combination searches
5'-(AC)x(TA)y-3'	(AC)(TA)	1	
3'-(TG)y(AT)-5'	(TA)(GT)		
5'-(AC)x(GA)y-3'	(AC)(GA)	7	
3'-(TG)y(CT)-5'	(TC)(GT)		
5'-(AT)x(CA)y-3'	(AT)(CA)	2	
3'-(TA)y(GT)-5'	(TG)(AT)		
5'-(AT)x(GA)y-3'	(AT)(GA)	2	
3'-(TA)y(CT)-5'	(TC)(AT)		
5'-(AG)x(CA)y-3'	(AG)(CA)	0	
3'-(TC)y(GT)-5'	(TC)(CT)		
5'-(AG)x(TA)y-3'	(AG)(TA)	0	
3'-(TC)y(AT)-5'	(TA)(CT)		
5'-(CA)x(TC)y-3'	(CA)(TC)	0	
3'-(GT)y(AG)-5'	(GA)(TG)		
5'-(CT)x(AC)y-3'	(CT)(AC)	7	
3'-(GA)y(TG)-5'	(GT)(AG)		
5'-(AG)x(GT)y-3'	(AG)(GT)	2	
3'-(TC)y(CA)-5'	(AC)(CT)		
5'-(TG)x(GA)y-3'	(TG)(GA)	3	
3'-(AC)y(CT)-5'	(TC)(CA)		
5'-(GT)x(TA)y-3'	(GT)(TA)	0	
3'-(CA)y(AT)-5'	(TA)(AC)		
5'-(AT)x(TG)y-3'	(AT)(TG)	1	
3'-(TA)y(AC)-5'	(CA)(AT)		

5

10

15

20

25

30

35

40

45

50

55

5'-(CT)x(TA)y-3'	(CT)x(TA)	0
3'-x(GA)y(AT)-5'	(TA)x(AG)	
5'-(AT)x(TC)y-3'	(AT)x(TC)	0
3'-x(TA)y(AG)-5'	(GA)x(AT)	
5'-(GT)x(TC)y-3'	(GT)x(TC)	1
3'-x(CA)y(AG)-5'	(GA)x(AC)	
5'-(CT)x(TG)y-3'	(CT)x(TG)	0
3'-x(GA)y(AC)-5'	(CA)x(AG)	
total occurrences		26

COMPOUND, NONPALINDROMIC, NO SHARED NTS BETWEEN REPEATS

double stranded locus

permutations 5'→3'

5'-(GA)x(CT)y-3'	(GA)x(CT)	0
3'-x(CT)y(AG)-5'	(AG)x(TC)	
5'-(CT)x(GA)y-3'	(CT)x(GA)	0
3'-x(GA)y(CT)-5'	(TC)x(AG)	
5'-(CA)x(GT)y-3'	(CA)x(GT)	0
3'-x(GT)y(CA)-5'	(AC)x(TG)	
5'-(GT)x(CA)y-3'	(GT)x(CA)	2
3'-x(CA)y(GT)-5'	(TG)x(AC)	
total occurrences		2

COMPOUND, PALINDROMIC, NO SHARED NTS BETWEEN REPEATS

double stranded locus

permutations 5'→3'

5'-(GA)x(TC)y-3'	(GA)x(TC)	0
5'-(AG)x(CT)y-3'	(AG)x(CT)	0
5'-(TC)x(GA)y-3'	(TC)x(GA)	0
5'-(CT)x(AG)y-3'	(CT)x(AG)	0
5'-(CA)x(TG)y-3'	(CA)x(TG)	0
5'-(AC)x(GT)y-3'	(AC)x(GT)	1
5'-(TG)x(CA)y-3'	(TG)x(CA)	0
5'-(GT)x(AC)y-3'	(GT)x(AC)	1
total occurrences		2

5

10

15

20

25

30

35

40

45

50

55

NOT COMPOUND: IMPERFECT SIMPLE REPEATS

double stranded locus	permutations 5'→3'
5'-(CA) _x (AC) _y -3'	(CA)(AC) 13
3'-x(GT) _x (TG) _y -5'	(GT)(TG) 18
5'-(AC) _x (CA) _y -3'	(AC)(CA) 4
3'-x(TG) _x (GT) _y -5'	(TG)(TG) 9
5'-(GA) _x (AG) _y -3'	(GA)(AG) 20
3'-x(CT) _x (TC) _y -5'	(CT)(TC) 64
5'-(AG) _x (GA) _y -3'	(AG)(GA)
3'-x(TC) _x (CT) _y -5'	(TC)(CT)
5'-(AT) _x (TA) _y -3'	(AT)(TA)
3'-x(TA) _x (AT) _y -5'	(TA)(AT)
	total occurrences

The complete GenBank DNA sequence databases (version 84.0) were searched using the FindPatterns search algorithm within the University of Wisconsin Genetics Computer Group sequence analysis package (version 7.3). The individual strands of the double stranded sequences shown in the first column were used individually as queries either against the entire GenBank database (all species combined) or against the separate subdatabases representing the indicated phylogenetic groupings. For each query, such as $(AC)_x(AT)_y$, x and y were each designated to be >6 ; that is, any hit in the database was required to carry at least 6 units of each of the two constituent dinucleotide repeats. The final column designates the number of matches to the respective query within an in-lab collection of cloned soybean SSR sequences (unpublished data), isolated as small inserts containing either $(AC)_x$ or $(AG)_y$ sequences.

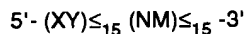
Although all 80 compound dinucleotide sequences have the potential to exist in a genome and to serve as targets for corresponding primers, only a small subset of these adjacent dinucleotide repeat combinations has been observed to occur at a reasonable frequency among compound SSRs represented within the DNA sequence databases and among SSRs cloned and sequenced from plant and animal genomes (see Table II). The great majority of these 80 total combinations occur at surprisingly low relative frequencies in eukaryotic genomes.

Nearly all of the compound repeats in this high frequency subset have perfect nucleotide periodicity whereby the two adjacent constituent dinucleotide repeats share a common nucleotide that retains a constant periodic spacing across the junction and the entire compound structure. These compound SSR sequences are designated by Applicants as "in-phase", and include repeats such as $(AT)_x(AG)_y$, $(AC)_x(TC)_y$, and $(AT)_x(GT)_y$ where x and y independently are ≥ 2 and can be multiples of 0.5. Thirty-two of the 80 possible dinucleotide combinations fall into this in-phase category, and these are listed in Table II. All the remaining compound repeats are termed, "out-of-phase". The 32 individual in-phase compound sequences (excluding CG and GC), however, represent only 16 unique nonredundant single-stranded sequences. The 2-fold redundancy derives from the possibility of positioning the repeating in-phase nucleotide as either the first or the second base in the core repeats (i.e., $(AC)_x(AT)_y$ and $(CA)_x(TA)_y$, although nonidentical as individual primers, represent the same compound sequence). In addition, these 16 canonical, nonredundant sequences represent the complementary strands of only 8 individual double-stranded compound repeat loci (see Table II). In other words, a compound dinucleotide repeat as a locus in double stranded DNA could be recognized by any of four different single stranded oligonucleotide primers, out of the total set of 32 possible permutations. For example, a locus $(AT)_6(AG)_6$, is a target site for the four hypothetical primers, $(AT)_x(AG)_y$, $(TA)_x(GA)_y$, $(CT)_x(AT)_y$, and $(TC)_x(TA)_y$ (with $x, y \leq 6$).

By chance alone, and assuming no nucleotide bias in a source genome, each of the 80 different perfect compound dinucleotide permutations would be expected to occur in the genome at equal frequencies. As mentioned above, however, it was already discovered by Applicants that the in-phase subset are more abundant compared to the out-of-phase set. Further, the two possible permutations of the constituent repeats for some of the in-phase compound sequence combinations appear to be represented at widely differing frequencies in a given genome. For example, the compound repeat, $(AT)_6(AG)_6$ appears to be at least 5 times more abundant in plant genomes than its permutant counterpart, $(AG)_6(AT)_6$; and, $(AC)_6(AG)_6$ is much more abundant in primate genomes than $(AG)_6(AC)_6$. Both from a systematic analysis of cloned sequence databases and from an empirical examination of both plant and animal genomes, these few, most frequently occurring compound dinucleotide repeats are known by Applicants and therefore are fully predictable. This knowledge serves to reduce to only a few the number of different compound SSR primers that will be successful for producing an adequate number of SSR-to-adaptor co-amplification products using the present invention.

Thus, of the 80 total compound dinucleotide sequence permutations that are possible by random arrangement of nucleotides, only a few (nearly all of which are in-phase) are present in plant and animal genomes at a measurable frequency, and only these few, therefore, are required to detect a large proportion of the compound SSR loci present in any given genome. Experimental data demonstrate, however, that specific primers representing these 16 compound sequences are not equally effective at recognizing and priming from the respective target locus. Such differences in primer efficiency were determined empirically to result from each constituent repeat's base composition (AT-richness), in combination with the relative position (5' anchoring versus 3'-priming) of each constituent repeat within the primer.

Table II also lists the dinucleotide permutations for which the spacing of a shared nucleotide is not preserved (termed, "out-of-phase") or for which no nucleotide is shared between the two constituent repeats. The latter category contains dinucleotide combinations that are both palindromic and nonpalindromic. Although not nearly as frequent in eukaryotic genomes as the in-phase sequences, some of the out-of-phase compound SSR sequences nonetheless appear to be present in most genomes. Therefore, primers that correspond to such out-of-phase repeats also are expected to serve as initiation sites for primer-extension from their respective target loci in the genome. Preferred anchored primers of the instant invention where primer nucleotide periodicity is not specifically designated may be defined by formula I for dinucleotide repeats:

Formula I

5

where

X = A, C, T, or G ; Y = A, C, T, or G ; X_Y
N = A, C, T, or G ; M = A, C, T, or G ; N_M

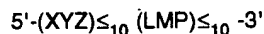
10

and where XY ≠ NM

Herein, these are abbreviated as (XY) #(NM)# and by formula II for trinucleotide repeats:

Formula II

15



where

20

X = A,C,T,or G; Y = A,C,T,or G; Z=A,C,T, or G
and X, Y, Z are not the same single base;

where

25

M = A,C,T,or G; N = A,C,T,or G; P=A,C,T,or G
and M, N, P are not the same single base;

and where XYZ ≠ NMP

30

Generally the primers of formulae I and II will consist of oligonucleotides of 10-60 nucleotides in length that contain two different, constituent simple sequence repeats that are directly adjacent to one another, with no intervening non-repeat nucleotides. From the 5' end, this oligonucleotide contains a simple sequence repeat of up to 15 repeat units in length, followed immediately 3' by a second simple sequence repeat that is also up to 15 repeat units in length.

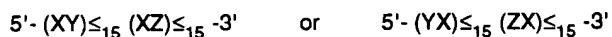
Preferred anchored primers of the instant invention where primer nucleotides are specifically designated to be in-phase may be defined by the formula III for dinucleotide repeats:

35

Formula III:

This formula describes a subset of sequences covered broadly by Formula I.

40



where

45

X = A,C,T,or G; Y = A,C,T,or G; Z =A, C,T,or G

but where Y ≠ X ; Z ≠ X ; and Y ≠ Z

Herein, these are abbreviated as (XY) #(XZ)# or (YX) #(ZX)#.

50

Typically oligonucleotides of formula III are 10-60 nucleotides in length and contain two different, constituent simple sequence repeats that are directly adjacent to one another, with no intervening nonrepeat nucleotides. From the 5' end, this oligonucleotide contains a simple sequence repeat of up to 15 repeat units in length, followed immediately 3' by a second simple sequence repeat that is also up to 15 repeats. Each repeat shares a common nucleotide, which is retained at a consistent periodicity across both constituent repeats, occupying either the first or the second position within each repeat.

55

In an especially preferred embodiment of Applicants' invention, PCR amplification to detect polymorphisms is carried out using restriction fragmented DNA modified with appropriate adaptors, wherein the primer pair used is comprised of one primer which is of the first type described above (Zabeau EP 534,858; an AFLP primer) and the second

primer is one of Applicants' unique perfect compound SSR primers described above.

One of skill in the art will also appreciate that Applicants' unique compound SSR primers can also be used in conjunction with a variety of other primer types which include for example, non-adaptor primers, primers of fixed sequence, arbitrary primers or any primer that might hybridize with currently known or unknown dispersed repeated sequences in the genome. An example particularly suited to the present invention would be where the other primer is

of completely arbitrary sequence such as a RAPD primer (Williams et al., *Nucleic Acids Res.* 18, 6531, (1990)). RAPD primers have been used to generate polymorphic markers from the amplification of genomic DNA. Preferably the nucleotide sequence of the RAPD primers would be about 9 to 10 bases in length, between 50 and 80% G+C in composition and contain no palindromic sequences. Amplifications using RAPD primers alone are typically done using short primers and low annealing temperatures which maximizes the probability that several randomly distributed loci on the genome will produce amplification products. Because the incidence of any particular RAPD binding site within the genome is relatively low, this methodology would serve to restrict the amount of the genome that is subject to amplification with any particular primer combination. It is contemplated that the selectivity of the RAPD methodology would serve an enrichment function, similar to the selection function provided by use of a biotinylated adaptor in the conventional AFLP method to enrich for only a subset of randomly distributed genomic regions, from which a manageable number of co-amplification could then occur.

PREPARATION OF GENOMIC DNA:

Restriction digested fragments:

Target DNA useful for amplification in the present invention was comprised of restriction fragments generated from Taq I + Pst I or Taq I + HindIII digestion of eukaryotic genomic DNA, further modified by the ligation of specific adaptor sequences. Genomic DNA was isolated from soybean and corn using either the CTAB/chloroform extraction and CsCl/centrifugation method of Murray and Thompson (Murray et al., *Nuc. Acid Res.*, 8, 4321, 1980) or a urea extraction miniprep procedure (Chen et al., *The Maize Handbook*, M. Freeling and V. Walbot, eds., (1993) pp 526-527, New York). Mammalian and salmon genomic DNAs were purchased from commercial sources (Sigma (St. Louis, MO); Clontech (Palo Alto, CA)). Genomic DNA was prepared for amplification reactions by complete restriction endonuclease digestion followed by ligation of site-specific double-stranded adaptors. Methods for this type of adaptor design and construction are well known in the art, and examples are given by Zabeau, EP 534,858.

It is preferred if a combination of two different restriction enzymes having 4-bp and 6-bp recognition sites, respectively, are used for the preparation of target DNA. Examples of suitable restriction enzymes are Taq I and Pst I however, any restriction enzymes having 4-, 5-, 6- or even 8-bp recognition sites also are appropriate providing their activities are not inhibited by target site DNA methylation or other nonmendelian mechanisms of selective nucleotide modification. Any combination of such restriction enzymes are potentially suitable. Other restriction enzymes suitable for the present invention may include but are not limited to the hexanucleotide site enzymes, EcoRI, DraI or BamHI, the tetranucleotide site enzymes, Sau 3AI, MboI, MseI, Tsp509I or AluI, the pentanucleotide site enzymes HinfI or AclI, and the octanucleotide site enzymes, PmeI, PacI, or SmaI.

Genomic DNA is digested with a first enzyme such as Taq I, followed by further digestion with a second enzyme such as Pst I, according to standard protocols, such as that given by Zabeau (EP 534,858). The digestions generated from each input genomic DNA are a mixture of symmetric fragments, bordered at both ends by either Taq I or Pst I sites, and asymmetric fragments, each flanked by a Taq I site and a Pst I site.

Adaptors:

Double stranded adaptors are generated by annealing the two partially complementary single stranded component oligonucleotides of each pair (examples are listed in Table I). Since restriction endonucleases cleave genomic DNA molecules at specific sites, amplification of restriction fragments can be achieved by first ligating synthetic oligonucleotide adaptors to the ends of restriction fragments, thus providing all restriction fragments with two common flanking tags which will serve as anchor bases for the primers used in PCR amplification. Typically, restriction enzymes either produce flush ends on a DNA fragment, such that the terminal nucleotides of both strands are base paired, or generate staggered ends in which one of the two strands protrudes to give a short (1-4 nt) single strand extension. In the case of restriction fragments with flush ends, adaptors are used with one flush end. In the case of restriction fragments with staggered ends, adaptors are used that have a single stranded extension complementary to the single stranded extension on the restriction fragment. Consequently, each type of restriction site end is specifically recognized by a particular adaptor by virtue of the complementarity of the matched ends. In addition, the DNA sequence of the entire length of each adaptor type differs from that of other adaptor types (see Table I). Typically, the adaptors used are comprised of synthetic single-stranded oligonucleotides which are in part complementary to each other, and which are usually

approximately 10 to 30 nucleotides long, preferably 12 to 22 nucleotides long, and which form double stranded structures when mixed together in solution. Using the enzyme T4 DNA ligase, the adaptors are joined covalently and specifically to the complementary ends of individual DNA molecules in the mixture of restriction fragments generated from a particular genomic DNA source. Using a large molar excess of adaptors over restriction fragments ensures that all restriction fragments will receive adaptors at both ends. These adaptors are not usually phosphorylated. These ligated adaptors then serve as templates for the adaptor-directed PCR primers.

In one embodiment of the invention, all restriction fragments from the genome carry the same adaptor at both ends, and a single PCR primer corresponding to that adaptor sequence can be used to amplify simultaneously from the fragments. The simultaneous amplification of several different restriction fragments is often referred to as multiplex PCR amplification. Since in such a case all restriction fragments are bordered at both ends by the same adaptor, it is obvious that primer extension and PCR amplification of a mixture of tagged restriction fragments will amplify all restriction fragments in a synchronous fashion. In another embodiment using two or more different restriction enzymes to cleave the DNA, two or more different adaptors are ligated to the ends of the restriction fragments. In this case, two different PCR primers, each matching the sequence of a particular adaptor, can be used for exponential amplification from a subset of the restriction fragments. In one preferred embodiment using two or more restriction enzymes, both adaptors are unmodified, and the fragment mixture is enriched using a pre-amplification step. In another preferred embodiment using two or more restriction enzymes, the adaptor corresponding to one of the restriction enzyme site ends is covalently linked to a biotin molecule. Using standard methods for isolating biotinylated molecules, this design allows for the selection, from a complex mixture of restriction fragments, of only those bordered on one or both ends by a biotinylated adaptor. Both of the two possible selection steps reduces the complexity of the starting mixture of restriction fragments and constitutes an enrichment step prior to the PCR amplification, thereby reducing in certain instances the background of fragments with same-site ends. In yet another embodiment one of the amplification primers may be radiolabeled for identification of the products via autoradiography, or may be modified with fluorescent tags for fluorescence detection of products. Methods of labeling nucleic acids and suitable labels are well known in the art (see Sambrook *supra*). For example a radioisotope suitable in the present invention is ^{33}P Phosphate, incorporated at the 5' end of one strand of the adaptor by a phosphate group transfer from of $[\gamma\text{-}^{33}\text{P}]\text{ATP}$ under kinasing conditions.

Double stranded adaptors (with or without biotin labels) are generated by annealing the two partially complementary single stranded component oligonucleotides of each pair (listed in Table I). For example, the double stranded Taq I adaptor (Taq-Ad) is produced by combining the single-stranded Taq. AdF and Taq. AdR oligonucleotides under favorable annealing conditions. To generate the biotinylated double stranded Pst I adaptor (biotin-Pst-Ad), the single stranded oligonucleotides biotin-Pst.AdF and Pst.AdR similarly are combined. If restriction enzymes other than those listed in Table I are used, then the adaptors must be designed to carry the appropriate protruding single-stranded-ends for a given restriction 1/2-site. In a preferred embodiment, each adaptor contains a single base alteration within the restriction half-site it carries, so that the reconstructed site generated by each ligation event cannot be re-digested. Therefore, the artisan will appreciate that it is possible that restriction digestion and ligation can be performed simultaneously under the appropriate buffer and temperature conditions.

DNA AMPLIFICATION:

Although the basic protocols for the amplification of nucleic acids are well known in this art, significant modifications of those protocols were necessary in order to achieve optimal amplification of DNA fragments and detection of polymorphic products. Several factors were found to significantly influence these amplifications, including variation in the thermocycling parameter of the PCR protocols, variation in the labeling of the primers, and the length and nucleotide composition of the primers.

Thermocycling variation in PCR:

The efficiency of amplification by both 5'-anchored simple and compound SSR primers was tested on soybean, corn, and mammalian templates, using thermocycling profiles having either constant temperature or touchdown annealing conditions. Both the adaptors and the SSR-directed primers in these amplifications were designed to have T_m 's within a relatively narrow range (38-45°C in 50 mM Na or K salt), so that any primer pair chosen for an amplification would have approximately the same optimal annealing temperature.

Three different constant annealing temperatures, 52°C, 58°C and 60°C were tested using a standard 3-step per cycle protocol. Although the results from each test varied from the others (the higher the temperature, the fewer the products generated), it is clear that the efficiency of primer discrimination at most target loci was found to be unacceptably inefficient at any of these constant annealing temperatures. Every "product" from these amplifications was represented by a small family of bands, and the products within each family differed by multiples of two; the length of each dinucleotide repeat. However, this "stuttering" effect and potential nonspecific product formation was minimized when

a touchdown thermocycling protocol (Don, et al., *Nuc Acids Res*, 19, 4008 (1991)) was used. In touchdown amplification, the annealing temperature begins deliberately high, then is incrementally lowered in successive cycles, down to a desired, "touchdown" annealing temperature. Touchdown temperatures of 59°C, 58°C and 56°C were tested for some primer combinations. For most SSR-directed primers, the 56°C final touchdown conditions produced the greatest number of specific, non-stuttering bands. For the purpose of the present invention, therefore, it is most preferred if nucleic acid amplification be conducted according to a touchdown protocol where 56°C is the optimal final annealing temperature. However, depending on the actual composition of the primers, a particular amplification may involve final annealing temperatures of 55°C-60°C.

Another variable in the thermocycling protocol that was explored is the method by which the amplification reactions are initiated. Typical PCR amplification protocols initiate with a cold start; all reagents necessary for DNA amplification are present in the reaction mixture prior to the first denaturation step. In contrast, a hot start protocol calls for the exclusion of one key reaction component, typically either the primer, nucleotides or polymerase, from the mixture during reaction setup and the first denaturation. This component then is added following denaturation, and primer extension can proceed.

The cold start protocol allows for the possibility that primers will anneal under nonstringent conditions both to template sites that are not necessarily a perfect match, and to multiple, staggered sites within a target locus. Often this method leads to a stuttering effect of the amplification products on the gel. In contrast a hot start protocol prevents spurious primer annealing to incorrect template sites at ambient temperatures prior to the first denaturation, and generates products that resolve more sharply and discretely on the gel. When otherwise identical amplification reactions were performed using the cold start and hot start protocols, it was found that for nearly every 5'-anchored simple SSR-directed primer, a cold start produced unacceptably indistinct products on the gel. Much sharper products were generated using hot start. In contrast, product resolution was found to be more consistent between cold and hot start methods when using compound SSR primers. In spite of the potential drawbacks in product resolution, cold start amplification was easier to perform, particularly when processing large numbers of samples, and was routinely found to be sufficient to generate amplification products that could be distinguished as polymorphic between genomes. Therefore, a slight gain in product resolution is sacrificed in a cold-start protocol in exchange for greater speed and ease of reaction setup. If 5'-anchored simple SSR primers are used, a hot-start is preferred, but a cold-start is adequate and sufficient for amplification reactions involving most compound SSR-directed primers.

Choice of primer labeling:

The complementary strands of a duplex DNA molecule usually resolve independently to slightly different positions on a denaturing polyacrylamide gel. If both strands of a DNA duplex are radiolabeled, then the autoradiograph will show a separate band representing each strand of each amplification product. Resolution of only a single band for each amplification reaction product on the denaturing polyacrylamide gel requires that only one strand of each product be labeled. To achieve this, only one of the two primers in any given pair used for an amplification reaction should be labeled. Either ³²P or ³³P can be used as radiolabels, although ³³P images are generally sharper on an autoradiograph. Alternatively, a variety of different fluorophores can be incorporated into the primer; the resulting products can then be detected using a fluorescence detection system.

The effects of radiolabeling the SSR primer, in comparison to the alternate labeling of the adaptor-directed primer, were explored. In all protocol variations, radiolabeling the adaptor-directed primer resulted in significant background. Generally, the lanes on the gel contained a few discrete bands, but the major result in every case was a smear of products distributed along the entire length of the lane. In contrast, when only the SSR-directed primer was labeled in the amplification reactions, the products were much more discrete on the autoradiograph and not associated with any significant lane background. This difference likely results from the high abundance of adaptor target sites on a large proportion of the template fragments (from which even linear amplifications with labeled primers collectively will lead to significant backgrounds). In contrast, the labelled SSR primers have far fewer target sites in the template mixture, and most become part of a productive, exponential amplification. Hence 5'-labeling of only the SSR primer is preferred, and would apply to labeling with either radioactive or fluorescent tags.

Variations in SSR and adaptor primer design:

Variability in the SSR-to-adaptor amplification reaction, and therefore in the products obtained, results not only from the reaction and thermocycle setup conditions described above but also from subtleties in the design of the primers used in these amplifications. Once a particular compound SSR has been chosen as the target locus sequence for this assay, then either partially or entirely different sets of amplification products can still be controlled by altering any one of the following primer design criteria:

- a) the number and base composition of the 3'-extension nucleotide(s) on the adaptor-directed primer;
 b) the relative lengths of the two constituent simple repeats that comprise the compound SSR primer;
 c) the particular strand of the double-stranded compound SSR locus chosen to correspond to the single-stranded primer (i.e., the directionality of the primer);
 5 d) choice of restriction enzymes and SSR targets for a particular genome.

a) Length of the adaptor-directed primer:

An adaptor-directed amplification primer which corresponds to the sequence of one of the synthetic adaptors
 10 ligated to the restricted ends of the genomic DNA can carry a variable number of arbitrary sequence nucleotides (zero to ten) at its 3'-end. These variable 3'-nucleotides on the primer anneal specifically to sequences that are directly adjacent to the adaptor and restriction site and whose sequences are not known, *a priori*, on any particular genomic restriction fragment. The recognition of each such primer to only a subset of all possible fragments in the template mixture provides exquisite specificity in the amplification reaction (Zabeau, EP 534,858). Such primers, otherwise
 15 identical in sequence except for differences in the few 3'-most nucleotide(s), can amplify completely nonoverlapping sets of amplification products and behave much like allele-specific amplification primers (Newton et al., (1989) *Nuc. Acids Res* 17: 2503; Kwok et al., (1990) *Nuc. Acids Res.* 18: 999; Wu et al., (1989) *Proc. Natl. Acad. Sci. USA* 86: 2757). The key difference, however, is that use of these adaptor-directed primers requires no prior sequence knowledge of the genomic locus to be amplified, and each primer will selectively co-recognize multiple target sites in a template
 20 DNA mixture.

In general, the longer the variable 3'-extension, the more selective or restrictive the primer. This 3'-extension contains arbitrary, nondegenerate or partially degenerate bases, which restrict annealing of the primer to only a subset of the total number of potential target sites, thus leading to a reduction in the real number of co-amplified products. The addition of each nondegenerate nucleotide onto the 3'-extension leads hypothetically to 4-fold greater template discrimination. In addition, different single nucleotides at the 3'-most base position(s) confer unique template specificities to otherwise identical primers. Thus, varying both the number and composition of the 3'-selective nucleotides on the
 25 adaptor-directed primer is sufficient to generate individual, either partially or completely, nonoverlapping sets of amplification products from the same template when paired with a given SSR-directed primer. The choice of which 3'-extension to use for a particular amplification is largely a matter of chance, but still will depend largely upon relative
 30 nucleotide frequencies in a target genome and upon the abundance in the genome of the specific SSR that serves as the other priming site.

b) Relative lengths of the two constituent simple repeats comprising the compound SSR primer:

Every simple and compound SSR locus in the genome is a double stranded structure whose individual strands
 35 carry different permutations of nucleotides. Unless the core dinucleotides of a compound repeat are palindromic (e.g., (CA)_x(TG)_y or (AG)_x(CT)_y), a single-stranded primer that may specifically anneal to one strand at a particular SSR locus will not anneal to the opposite strand. None of the in-phase compound SSRs is palindromic and only 8 of the 90 possible dinucleotide permutations represents such a palindrome. Therefore, all in-phase and most out-of-phase SSR-
 40 directed primers will primer-extend from each genomic target locus in a polar, unidirectional manner, and any compound SSR locus can be recognized and amplified from by any of four primer classes. For example, the compound in-phase SSR locus,

45 5' -CACACACACACACACACACATATATATATATATATATA-3' SEQ ID NO.:1

3' -GTGTGTGTGTGTGTGTGTGTATATATATATATATATAT-5', SEQ ID NO.:2

50 can be recognized by four different canonical primer classes:

5'-(AC)_x(AT)_y-3',

5'-(CA)_x(TA)_y-3',

5'-(AT)_x(GT)_y-3',

and 5'-(TA)_x(TG)_y-3', where the 5'-most repeat in each serves primarily to anchor the primer to the template, and
 55 the 3'-most repeat serves a primer-extension function (see Figure 1c).

Each of these four canonical primer classes can include a wide range of individual primers, all differing by the length of the two constituent repeats within the primer. Changes in the lengths of these constituent repeats have profound effects on primer efficacy and the fidelity of reproducible amplifications. In general, the longer the 5'-anchoring

repeat (i.e., the value of x, above) relative to that of the 3'-priming repeat (the value of y), the better the primer's specificity and priming efficiency in the amplification. In addition, only a primer with a short 3' repeat will allow amplification from compound SSR loci containing very short downstream repeats.
compound SSR loci containing very short downstream repeats.

5

c) Polarity of the single stranded compound SSR-directed primer:

The choice of which strand of a double-stranded compound SSR locus to use as a primer can be extremely critical for determining the success of the SSR-to-adaptor amplification reaction. It should be noted that the only type of (AT)_n-containing primer that will lead to efficiently generated amplification products under standard conditions is one in which the (AT)_n sequence is very short (1.5-3 repeat units) and is situated as the 3'-primer extension end. An (AT)_n repeat of any length at the 5' end is completely inefficient as an anchor, and results in little or no amplification from a complex genomic template mixture.

d) Restriction site frequencies, nucleotide bias, SSR frequencies, and primer design considerations:

Ligation products carrying biotinylated adaptors may be selected out of each digestion/ligation mixture using a streptavidin or avidin coated support such as paramagnetic beads, as provided by Dynal Inc., (Lake Success, NY). This selected DNA does not have to be purified further from the beads for the subsequent amplifications. In one embodiment where two restriction enzymes are used and only the adaptor corresponding to the restriction enzyme with the hexanucleotide site is biotinylated, the selected DNA is a mixture of fragments bordered only by one or the other restriction site or flanked at each end by different sites. For example, if Taq I and Pst I are used, then only the Pst I adaptor is biotinylated. Following biotin selection, the Taq I-Pst I and Pst I-Pst I fragments are predicted to be present in the enriched fragment mixture at an approximate ratio of 30:1, respectively. All Taq I-Taq I fragments are effectively discarded. Methods for such a calculation will be apparent to one skilled in the art, for example: if the frequency of each nucleotide is known for the specific genome, then the symmetric and asymmetric restriction fragments will be present in the digestion mixture at predictable proportions. In general, a calculation can be made that derives from the following assumptions:

First, recognition sites for each restriction enzyme are present in the genome at differing absolute frequencies, which are a function of the number of nucleotides in the site and of the genome's nucleotide composition. Second, these absolute frequencies can be converted to relative frequencies, p and q, since the sum of the relative frequencies (p+q) is always equal to 1. For example (considering equal nucleotide frequencies and random nucleotide distribution in a genome):

35

site	absolute frequency	relative frequency
Taq I	$(0.25)^4 = 3.9 \times 10^{-3}$	$p = 0.9412$
Pst I	$(0.25)^6 = 2.44 \times 10^{-4}$	$q = 0.0589$

Finally, the relative frequencies of restriction fragments bordered by these sites are simply the products of the relative site frequencies for each fragment type. Therefore:

45

fragment	relative frequency
Taq I-Taq I	$p^2 = 0.8862$
Pst I-Pst I	$q^2 = 0.0035$
Taq I-Pst I and Pst I-Taq I	$2pq = 0.1109$

Therefore in this embodiment, utilizing restriction enzymes with 4- and 6-bp recognition sites and assuming no nucleotide bias, the biotin-selected DNA fragments represent only 11.1% of the genome. Different restriction enzymes with different site frequencies will lead to a greater or lesser proportion of the genome represented in the mixture of selected fragments. Using several restriction enzyme combinations will ensure better coverage of the genome than just a single enzyme combination.

The selected DNA fragments may be used as a pooled template mixture for polymerase chain reaction amplifications using one each of a primer corresponding to one of the adaptors (the one that was not biotin-selected) and a primer directed to a particular 5'-anchored simple or compound SSR sequence. One skilled in the art will appreciate that a detectable product will result from any single genomic template fragment only when exponential amplification occurs between the adaptor-directed primer and an oppositely oriented SSR-directed primer (see Figure 1a). Multiple

amplification products are expected from each template DNA mixture since the SSR and adaptor sequences are not single-copy sites. The multiplex ratio (the number of co-amplified products) of each amplification reaction is affected by the absolute genomic copy number of a specific SSR sequence, and can be adjusted experimentally for a given SSR by altering either the level of degeneracy of a simple SSR primer's 5'-anchor or the number and quality of the nondegenerate selective nucleotides at the 3' end of the adaptor-directed primer. For example, assuming equal frequencies for all four nucleotides in the genome, the addition of each successive nondegenerate nucleotide onto the 3'-end of the adaptor-directed primer leads to a 4-fold reduction in the number of co-amplified products from a given template mixture (see Figure 1b). Therefore, an adaptor primer carrying zero selective nucleotides at its 3' end (e.g., Taq.AdF; see Table I) will co-amplify 4 times as many templates as will a primer with a single, nondegenerate 3'-nucleotide (e.g., Taq.pr6, whose 3'-extension is -A). Similarly, this primer will co-amplify 4 times as many template fragments as a primer carrying 2 selective nucleotides, and so on. In general, the degree of selectivity of the adaptor-directed primer can be estimated using the formula, $1/4^{2n}$, where n=the number of selective bases. It should be cautioned that although convenient, this simplified calculation does not take into account the base composition of the genome, nor of the recognition sites of the restriction endonucleases used to produce the genomic fragments.

1. Although most DNA fragments in the bead-selected or pre-amplified mixture should be bordered at one end by the adaptor corresponding to the adaptor-directed primer to be used in the PCR, only a subset of these fragments are expected to carry an internal simple sequence repeat region complementary to a particular SSR primer. Thus, amplification products will be generated and detected only from the subset of target molecules that not only are flanked by the primer-specific adaptor but also contain an internal repeat sequence matching the SSR primer. It should be noted that absolute frequencies for the different repeats can vary widely within a species and are not accurately known for most plant genomes. Preliminary studies indicate that in soybean, (AT)_n is at least twice as abundant as (CT)_n, which in turn appears to be somewhat more frequent than (CA)_n (Morgante & Olivieri, *Plant J* 3: 175 (1992); Akkaya et al., (1992) *Genetics*, 132:1131). In general, it is estimated that one SSR longer than 20bp exists in plant genomes once every 23-29kb, compared to a figure of 6kb in mammals (Wang et al., *Theor. Applied Genetics*: 88, 1 (1994); Morgante & Olivieri, *Plant J* (1993); Beckmann & Weber *Genomics* 12:627 (1992)). Frequencies of compound SSR sequences, however, have not been documented in the literature.

A completely degenerate 5'-anchor on a simple SSR primer should prime from every locus in the genome that carries that particular SSR sequence. Any degree of nondegeneracy introduced into the anchor will reduce the potential number of genomic target sites, and therefore the number of amplified fragments. In a genome with no nucleotide bias, the complexity of the co-amplified products is reduced by a factor of 4 for every anchor position that is assigned a nondegenerate nucleotide. Each self-anchoring compound SSR primer is expected to anneal and prime from every matching compound SSR locus in the biotin-selected fragment mixture (providing each target SSR locus has a sufficient length to allow complete hybridization by the primer).

Detection of polymorphisms between phenotypically related individuals:

Individual gel banding pattern differences of the co-amplified fragments between different templates (i.e., different genomes) indicate polymorphisms between the source genomes. The amplification products generated with any of the compound SSR-directed primers are a mixture of polymorphic and nonpolymorphic fragments. Compared to a conventional AFLP reaction (EP 0534858), from which most of the polymorphisms detected are dominant, Applicants' compound SSR-to-adaptor multiplexed amplification method generates a greater proportion of codominant polymorphisms. Although many of the amplification products generated by this scheme are nonpolymorphic between closely related strains, the proportion of polymorphic products increases between more distantly related lines. In general, genomic polymorphisms can be detected using the SSR-to-adaptor multiplexed amplification among individuals from within a species as well as between species; the greater the evolutionary distance between the genomes being compared, the more polymorphisms expected. Both dominant and codominant polymorphisms can be detected. In either case, a polymorphism revealed by this method may result from any one or a combination of possible causes:

- 1) One or both restriction sites bordering a given genomic region are missing in one genome (analogous to an RFLP, but here detected by a different method). This may be visible either as a dominant or a codominant difference between genomes;
- 2) Insertion or deletion differences exist between genomes, within the genomic fragment bordered by common restriction sites (this should be visible as a codominant polymorphism providing the amplification distance is not too great for either allelic fragment);
- 3) Length differences in the simple sequence repeat between genomes can lead to codominant polymorphic amplification products, generally differing in length by multiples of the repeat unit;
- 4) Single base differences are present between genomes in the region immediately adjacent to the restriction site, such that the 3'-selective portion of the adaptor-directed primer can discriminate between dissimilar templates, in

a manner analogous to an allele-specific amplification.

Because of all these potential sources of polymorphism, the information content on a per locus basis for this type of multiplexed amplification assay is very high. A simple estimate can be made for the minimum number of nucleotide positions at a locus that are informative (i.e., at which a polymorphism may be detected). For templates digested with both a 4- and 6-bp cutter:

4 (within the 4 bp restriction site)
 + 6 (within the 6 bp restriction site)
 + 0 to 10 (sequence immediately adjacent to the adaptor-specific restriction site)
 + 0 to 30 (number of nucleotides within the SSR assayable for length variability between genomes)
 = 10 to 50 nucleotides per amplified locus may be informative for producing a polymorphism between individual genomes. The first three factors in this sum result from single nucleotide variation (e.g., substitutions) between genomes, whereas the fourth factor in the sum results from repeat length variation. Although small insertions and deletions distributed in the entire genome can contribute to the detection of length variability in this as well as other genome assays, the greatly increased probability for repeat length variation at each SSR target locus results in an "above background" level of length polymorphism detectable in the products.

In comparison, the information content for RFLP is 8-12, for RAPDs is 16-18, and for a standard AFLP assay is generally 10-15 nucleotides. Furthermore, compared to conventional AFLP and RAPD technologies, a larger proportion of the polymorphism assayable by this SSR-to-adaptor amplification method is detectable as codominant differences between genomes.

SSR-based polymorphisms detected using this SSR-to-adaptor, or SAMPL, amplification method can be converted into more conventional and convenient single-locus SSR markers. This conversion can be performed, for example, if the multiplexed approach is used to quickly screen through hundreds or thousands of possible polymorphisms between genomes, and if it then is desirable to subsequently assay a chosen subset of these polymorphisms either at a larger, more high throughput scale or in order to examine polymorphism at these particular loci more quickly and nonisotopically. This conversion process requires that the desired band be excised from the SAMPL gel and then sequenced. From the nucleotide sequence deduced for the unique sequence flanking the SSR, a "locus-specific" primer can be designed, which flanks and is oppositely oriented towards the SSR. This unique primer can then be paired with a general adaptor-directed primer and used to amplify from the original fragment mixture. The resulting adaptor-to-unique primer PCR product then can be sequenced to discover the other unique flanking sequence of the SSR, and the second locus-specific primer can be designed. Finally, the two oppositely oriented locus-specific flanking primers are used as a pair to amplify the region spanning the desired SSR locus in a target genome.

EXAMPLES

MATERIALS AND METHODS

Restriction enzymes, ligases and polymerases used in the following examples were obtained from BRL Life Technologies (Gaithersburg, MD) or New England Biolabs (Beverly, MA).

The source of the soybean cultivars, Bonus and soja PI 81762, was Theodore Hymowitz, University of Illinois. All other soybean lines, including the *G. max* cultivars wolverine, NOIR-1, N85-2176, Harrow, CNS, Manchu, Mandarin, Mukden, Richland, Roanoke, Tokyo and PI 54-60, and the *G. soja* accession PI 440-913, were obtained from the USDA Soybean Germplasm Collection, University of Illinois (Dept. Agronomy, Turner Hall, Urbana, IL). The source of the *Z. mays* inbred cultivars, B73 and Mo17, and the elite lines, LH82, LH119 and LH204, was Holden's Foundation Seeds, Williamsburg, IA. The source of the *Z. mays* inbred line, CM37, was Benjamin Burr, Brookhaven National Laboratory, Upton, NY. The AEC272 and ASKC28 *Z. mays* lines were obtained from Dr. Denton Alexander, University of Illinois. Genomic DNA from five different human sources, as well as from salmon and mouse BABL/c, was purchased from commercial sources (Sigma, St. Louis, MO or Clontech, Palo Alto, CA).

Reagents, buffers and protocols used for restriction digests, ligations, 5'-end phosphorylation-labeling of primers and PCR amplifications are given below in Table III.

TABLE III
REACTION PROTOCOLS
TEMPLATE PREPARATION AND AMPLIFICATION REACTION SETUP

<u>Restriction Digestion</u>				
<u>Reagent</u>	<u>Stock</u>	<u>Final Conc</u>	<u>Amount Stock Used</u>	<u>Comments</u>
Genomic DNA	High-quality	.0833 ug/ul	2.5 ug	should be of the highest quality
10X Buffer	10X	1X	5 ul	compatible with most restriction endonucleases and with the subsequent ligation reaction
	100 mM Tris-acetate	10 mM		
	100 mM Mg-acetate	10 mM		
	500 mM K-acetate	50 mM		
	50 mM DTT	5 mM		
	pH 7.5			
<u>Restriction enzymes</u>				
Taq I (4 bp recognition site)	10 units/ul	5 units/ug DNA	1.25 ul	neither should be methyl-sensitive 4 bp recognition site
Pst I (6 bp recognition site)	10 units/ul	5 units/ug DNA	1.25 ul	6 bp recognition site
H ₂ O			bring to 50 ul	
<u>Adaptor ligation</u>				
<u>Reagent</u>	<u>Stock</u>	<u>Final Conc</u>	<u>Amount Stock Used</u>	<u>Comments</u>
Taq I-Ad	50 pmol/ul	.833 pmol/ul	1.0 ul	double strand adaptor for 4 bp enzyme
Biotin-Pst I-Ad	5 pmol/ul	.0833 pmol/ul	1.0 ul	5'-biotinylated adaptor for 6 bp enzyme
10X Buffer		1X	1.0 ul	same buffer as for restriction digests
ATP	10 mM	0.2 mM	1.2 ul	

5
10
15
20
25
30
35
40
45
50
55

T4 DNA Ligase	1 unit/ul	1 unit/reaction	1.0 ul	
H ₂ O			4.8 ul	bring to 10 ul total
<u>Primer radiolabeling</u>				
<u>Reagent</u>		<u>Stock</u>	<u>Final Conc</u>	<u>Amount Stock Used</u>
SSR primer	50 ng/ul		5 ng/ul	3.0 ul
10X Kinase Buffer	10X 600mM Tris-Cl		1X 60mM	3.0 ul
	100mM MgCl ₂ 150 mM DTT pH 7.8		10mM 15mM	
[gamma-33P]ATP (3000 Ci/mmol)	10 uCi/ul		1.67 uCi/ul	5.0 ul
				At 3000 Ci/mmol, equal to ~17 pmol. 32P also can be used, but with poorer gel resolution.
T4 polynucleotide kinase	10 units/ul		0.17 units/ul	0.5 ul
H ₂ O				bring to 30 ul
<u>PCR amplification</u>				
<u>Reagent</u>		<u>Stock</u>	<u>Final Conc</u>	<u>Amount Stock Used</u>
biotin-streptavidin selected template DNA				2 ul
Taq I adaptor directed-primer SSR-primer	50 ng/ul		1.5 ng/ul	0.6 ul (30 ng)
33P-labeled SSR primer	5 ng/ul		0.25 ng/ul	1.0 ul (5 ng)
unlabeled SSR primer	50 ng/ul		1.25 ng/ul	0.5 ul (25 ng)
				30 ng total (labeled + unlabeled)
				2.5 ug input genomic DNA in a final 200 ul selected volume is sufficient for 100 PCR amplifications eg., Taq.pr1, Taq.pr2, Taq.pr3, etc.

31

Preparation of genomic template DNA mixtures:

Genomic DNA was isolated from soybean (*Glycine max*) cultivars using a CTAB/chloroform extraction and CsCl/centrifugation method (Murry et al., *Nuc Acids Res*, 8, 4321, 1980), and from corn (*Zea mays*) cultivars using a urea extraction miniprep method (Chen et al., in *The Maize Handbook*, M. Freeling and V. Walbot, eds., (1993) pp 526-527 New York).

Purified genomic DNA was prepared for amplification reactions in a manner similar to that described by Zabeau (EP 534,858), by complete restriction endonuclease digestion followed by or coupled with ligation of site-specific double-stranded adaptors. The restriction enzyme combinations used for the following examples were either Taq I + Pst I or Taq I + Hind III (a combination of enzymes with tetra- and hexa-nucleotide recognition sites, respectively). Between 1 and 2.5 ug of high molecular weight genomic DNA was digested with 5 units/ug of Taq I in a 50 ul volume at 65°C for approximately 3 hours in a buffer containing 10 mM Tris acetate, 10 mM magnesium acetate, 50 mM potassium acetate, 5 mM dithiothreitol, pH 7.5, then digested further in the same buffer with 5 units/ug of Pst I or Hind III at 37°C for 3 h (Table III). The digestion products generated from each input genomic DNA were a mixture of symmetric fragments, bordered at both ends by either Taq I or Pst I (or Hind III) sites, and asymmetric fragments flanked by both a Taq I site and a hexanucleotide site.

Double stranded adaptors were generated by slowly annealing equimolar amounts of the two partially complementary single stranded component oligonucleotides of each pair (see Table III). The double stranded Taq I adaptor (Taq-Ad) at 50 pmole/uL was produced by combining 5000 pmole each of Taq.AdF and Taq.AdR single-stranded oligonucleotides with H₂O to a final volume of 100 uL. For the 5 pmole/ul Pst I and Hind III adaptors (biotin-Pst-Ad or biotin-Hind-Ad), 500 pmole each of the corresponding single stranded oligonucleotides for each were combined in a final volume of 100 uL. To generate the double-stranded molecules, all mixtures were incubated at sequentially decreasing temperatures: 65°C for 15 min, 37°C for 15 min, room temperature for 15 min, then finally at 4°C.

This section describes the method that utilizes biotin-streptavidin selection for enriching the genomic fragment mixture prior to the SSR-to-adaptor amplification. To each completed double digestion was added 10 uL of a mixture containing 1 unit T4 DNA ligase, 0.2 mM ATP, and the two double-stranded adaptors, Taq-Ad (50 pmole) and biotin-Pst-Ad or biotin-Hind-Ad (5 pmole), each carrying a different synthetic DNA sequence along its length and each complementary at one end to the single-stranded tetranucleotide or hexanucleotide overhangs on the genomic fragments (Table III). These adaptor ligation reactions were incubated at 37°C for 3 h. Each adaptor contained a single base alteration within the half-restriction site it carries, so that the reconstructed site generated by each ligation cannot be re-digested. Therefore, restriction enzyme digestions and ligations can be performed simultaneously providing the activities of all enzymes used share a common optimum reaction temperature.

A subset of ligation products, all carrying a biotinylated Pst I adaptor at least one end, was selected out of each digestion/ligation mixture using streptavidin coated paramagnetic beads (DynaL, Lake Success, NY). For each selection, 10 uL beads, washed once in 200 uL STEX (100 mM NaCl, 10 mM Tris-HCl, 1 mM EDTA, 0.1% Triton X-100, pH 8.0) then resuspended in 150 uL STEX, was added to each ligation reaction, and the mixtures incubated for one hour at room temperature on a gently rocking platform. The DNA-adhered beads then were selected out of each mixture using a magnetic rack support; the supernatant was aspirated away, and the beads were resuspended in 200 uL STEX. Four additional cycles of bead selection, washing and aspiration were performed. The final resuspension was transferred to a fresh tube, and the DNA-adhered beads selected in this final cycle were resuspended in 10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0 (100 uL for 1 ug input DNA or 200 uL for 2-2.5 ug input DNA). This selected DNA, which did not have to be purified further from the beads, was a mixture of Taq I-Pst I (or Taq I-Hind III) and Pst I-Pst I (or Hind III-HindIII) fragments, present at an approximate ratio of 30:1, respectively. The selected DNA fragments were used as pooled template for polymerase chain reaction amplifications using one each of a Taq I adaptor-directed primer and a simple sequence repeat (SSR)-directed primer.

For the alternative method of enriching the restriction fragment mixture prior to the final PCR amplification, both restriction site specific adaptors may be either biotin-modified or unmodified. Following adaptor ligation to the restriction fragments, the entire mixture is subjected to up to 16 individual amplification reactions, each of which employs a pair of individual adaptor-directed primers. One primer of each pair corresponds to one of the adaptor sequences, and the second primer to the second adaptor sequence. In each case, the amplification primer carries a single, randomly chosen nucleotide at its 3'-most end. Each primer pair will specifically amplify an approximately 1/16th subset of the original genomic fragment mixture. Any or all of the pre-amplified product mixtures derived from a common restriction fragment population then can serve as an enriched template for the subsequent amplification between an SSR-directed primer and primer representing either of the two flanking adaptors. In this case, this adaptor primer typically carries 2, 3, or more arbitrary 3'-nucleotides, with the nucleotide closest to the I restriction site perfectly matching the one nucleotide used on the pre-amplification adaptor primer.

EXAMPLE 1**AMPLIFICATION USING 5'-ANCHORED SIMPLE SSR PRIMERS IN A MULTIPLEXED SSR-TO-ADAPTOR AMPLIFICATION TO DETECT POLYMORPHISMS AMONG SOYBEAN CULTIVARS**

Example 1 describes the use of primers corresponding to simple SSR sequences flanked at the 5'-end by degenerate nucleotides, for the amplification of adaptor tagged restriction fragments and the subsequent detection of genetic polymorphisms among soybean cultivars. Adaptor-directed primers used in these amplifications are shown in Table I. The SSR primers, herein termed 5'-anchored simple SSR primers, are listed in Table IV.

TABLE IV
5'-ANCHORED SIMPLE SSR PRIMERS

Primer Name	Sequence(5'→3')/SEQ ID NO.	Length	T _m	
			(50mM Salt)	%GC
HBH(AG)8.5	HBHAGAGAGAGAGAGAGGA/SEQ ID NO.:3	20	45.7°C	45.00
BHB(GA)8.5	BHBGAGAGAGAGAGAGAG/SEQ ID NO.:4	20	47.7°C	52.50
DVD(TC)8.5	DVDCTCTCTCTCTCTCTCT/SEQ ID NO.:5	20	45.7°C	47.50
VDV(CT)8.5	VDVCTCTCTCTCTCTCTCTC/SEQ ID NO.:6	20	47.7°C	52.50
DBD(AC)7.5	DBDACACACACACACACA/SEQ ID NO.:7	18	41.8°C	47.20
BDB(CA)7.5	BDBCACACACACACACAC/SEQ ID NO.:8	18	44.1°C	52.80
HVH(TG)7.5	HVHTGTGTGTGTGTGTGT/SEQ ID NO.:9	18	41.8°C	47.20
VHV(GT)7.5	VHVGTGTGTGTGTGTGTG/SEQ ID NO.:10	18	44.1°C	52.80
CCGG(T)10	CCGGTTTTTTTTTT/SEQ ID NO.:11	14	23.4°C	28.60
GCGC(A)10	GCGCAAAAAAAAAA/SEQ ID NO.:12	14	23.4°C	28.60
BDBD(AC)6.5	BDBDACACACACACACA/SEQ ID NO.:13	17	39.5°C	47.00
BHBH(AG)6.5	BHBHAGAGAGAGAGAGA/SEQ ID NO.:14	17	39.5°C	47.00
VHVH(TG)6.5	VHVHTGTGTGTGTGTGT/SEQ ID NO.:15	17	39.5°C	47.00
CGG(CA)6.5	CGGCACACACACACACA/SEQ ID NO.:16	17	44.3°C	58.80

Adaptor modified, biotin-selected genomic DNA from the *Glycine max* strains, NOIR-1, N85-2176 and wolverine, and the *Glycine soja* cultivar, PI 81762, were prepared as described in the MATERIALS AND METHODS. SSR primers were 5'-end labeled with ³³P by combining fifty microcuries of [γ -³³P]ATP (New England Nuclear) with 150 ng of primer and 5 units of T4 polynucleotide kinase in a 30 μ L reaction (Table III). After incubation at 37°C for 1 h, a 1 μ L aliquot of this labeled primer mixture (containing 5 ng primer) was used directly in each amplification reaction.

All amplification reactions were performed simultaneously by cold start initiation and were set up together at room temperature, using a series of reaction cocktails. The master mixture, containing the four components common to all reactions, consisted of (per reaction) 2.0 μ L of 10X PEC buffer, 0.8 μ L all four dNTPs (5mM each), 13.0 μ L H₂O, and 0.1 μ L (0.5 units) Amplitaq DNA polymerase (Perkin Elmer Roche) (see Table III and MATERIALS AND METHODS). An aliquot of this master mixture was then added to the appropriate primers to make individual full-primer cocktails; for each final amplification reaction, 15.9 μ L of master mix was combined with 0.6 μ L unlabeled Taq I adaptor-directed primer (stock is 50 ng/ μ L), 0.5 μ L unlabeled SSR primer, and 1.0 μ L (at 5 ng/ μ L) ³³P-labeled SSR primer. The appropriate biotin-streptavidin selected template DNA (2 μ L each) was distributed into 0.2 mL microamplification tubes (Robbins Scientific, Mountain View, CA) and placed in individual wells of a Perkin-Elmer 9600 multiwell plate. Eighteen μ L of the appropriate full-primer cocktail then was added, to give all reactions a final volume of 20 μ L. This final combination of reaction components was quickly completed and the amplification reactions initiated on a Perkin Elmer 9600 thermocycler with as little delay as possible using either a constant 58°C annealing or a 58°C touchdown thermocycling

protocol:

5

10

15

20

25

Touchdown annealing	
denature	94°C, 3 min
1 cycle	94°C, 30 sec
	65°C, 30 sec
	72°C, 1 min
11 cycles	94°C, 30 sec
	64.4°C, 30 sec for 1st cycle, then decrease by 0.6°C per cycle for the next 10 cycles to a final 58.4°C
	72°C, 1 min
23 cycles	94°C, 30 sec
	58°C, 30 sec
	72°C, 1 min
Constant temperature annealing	
denature	94°C, 3 min
35 cycles	94°C, 30 sec
	58°C, 30 sec
	72°C, 1 min

30

35

40

45

50

55

The completed amplification reactions were diluted with an equal volume of formamide stop solution (98% deionized formamide, 2 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol), heated to 94°C for 3 min, then quickly chilled on ice. 2.5 μ L of each was immediately loaded onto a 4.5% or a 6% denaturing 30 x 40 x 0.4 cm polyacrylamide gel (7M urea, 4.5% acrylamide: N,N-methylene bis-acrylamide [19:1], 100mM Tris-HCl, 80mM boric acid, 1 mM EDTA, pH 8.3) that was first pre-run in the same tris-borate-EDTA buffer at 55W for 30 minutes. The loaded samples were electrophoresed at 55W (corresponds to -1400-1500V, 35-40mA) for two hours, then the gel was transferred to chromatography paper, vacuum dried for 2 h at 80°C, and exposed to Hyperfilm-MP (Amersham) or Biomax (Kodak) X-ray film with an intensifying screen at -70°C for 2 to 7 days.

Figure 2 shows a comparison of the amplification products using the 5'-anchored simple SSR-directed primers, DBD(AC)_{7.5}, HBH(AG)_{8.5}, and HVH(TG)_{7.5}. In all cases, the SSR primer was ³³P-labeled and used in combination with each of two different Taq I adaptor-directed primers, Taq.Ad.F and Taq.pr6, which carry zero and one 3' selective nucleotide, respectively (see Table I). Panel a shows the amplification products generated using constant temperature thermocycling and panel b shows the products from a touchdown protocol. Several different 5'-anchored simple SSR-directed primers have been tested, all of which are listed in Table IV. In general, it was found that all such simple SSR-to-adaptor amplifications required a 0-2 nucleotide extension on the 3' end of the adaptor-directed primer to give a suitable number of co-amplified products.

Regardless of the annealing temperature, the constant temperature annealing protocol produced bands on the gels that were extremely smeared and indistinct for nearly every primer combination tested. Raising the annealing temperature from 56° to 58°-59°C resulted in somewhat less smeariness (Figure 2a); the products of individual loci generally are discernable. However, the products are not discretely sized, and instead showed a high degree of "stutter" on the gel. The highest constant annealing temperature tested, 60°C, produced relatively few bands for some primer combinations, and no bands for others (data not shown). These results indicate that the efficiency of primer discrimination at most target loci is relatively inefficient when the annealing temperature is held constant throughout the thermocycling. Either the primer does anneal, but at multiple positions within a target locus (producing a stuttering effect), or it does not anneal stably to generate a product. Although an optimal constant annealing temperature for any primer pair ultimately should be determined empirically, it is likely that heterogeneously sized amplification products still will result using this thermocycling method.

In contrast, thermocycling using touchdown conditions (Don., et al., *Nuc. Acid Res.*, 19, 4008, (1991) are designed to lead to highly efficient target locus discrimination and to minimize or eliminate spurious priming by either of the

primers in the amplification. In touchdown amplification, the annealing temperature begins deliberately high, then is incrementally lowered in successive cycles, down to a desired, "touchdown" annealing temperature. Touchdown temperatures of both 59°C and 56°C were tested. For most SSR-directed primers, this range of final touchdown temperatures was optimal for producing a large number of relatively discrete bands on the gels (see Figure 2b for an example of the products resulting from 58°C touchdown reactions; other data not shown), and these bands were reproducible from experiment to experiment.

Individual banding pattern differences in the co-amplified fragments between different templates (i.e., different genomes) indicate polymorphisms between the source genomes. The majority of the amplification products generated by this scheme appear to be nonpolymorphic between the two closely related *G. max* strains, NOIR-1 and N85-2176; however, a greater number of polymorphic products are seen between the more distantly related *G. max*, wolverine and *G. soja* PI 81762 cultivars. In addition, some polymorphisms appear to be dominant (a band amplified from one genome, but no apparent corresponding band from the other), and a few are potentially codominant (bands of similar but nonidentical size amplified from each genome).

This analysis illustrates two important features of the ability to fine tune and customize this assay. First, the two different 5'-anchored simple SSR directed primers, HBH(AG)_{8.5} and DBD(AC)_{7.5}, when used against a common set of genomic templates, produced completely different amplification patterns. That these patterns reflect distinctly different subsets of amplification products is consistent with the idea (see Figure 1) that different subsets of restriction fragments from the genome are likely to carry different SSR target sites. Each band of the gel is the result of a productive amplification between a particular SSR sequence oppositely oriented relative to a Taq I site. Given the estimation for relatively large spacing between SSR sequences of all types in the soybean genome (Wang et al., *Theor. Applied Genetics*, 88, 1 (1994); Morgante & Olivieri, *Plant J* 3, 175 (1993)), it is unlikely that the SSR-to-Taq I site products derived from the HBH(AG)_{8.5} and DBD(AC)_{7.5} primers in this example cover any genomic loci in common.

A second feature illustrated by this example is that for any chosen SSR primer, the fewer the number of 3'-selective nucleotides on the adaptor-directed primer, the greater the number of co-amplified bands. In general, it is expected that the mixture of products generated using the n=0 version of the adaptor primer is more complex than that for an n=1 primer, and this product mixture is more complex than that for an n=2 primer, and so on. In this example, the Taq I adaptor-directed primer, Taq.AdF carrying zero 3'-selective nucleotides, consistently produced a greater number of labeled reaction products when paired with a given SSR primer, compared to primers carrying one selective nucleotide (Taq.pr6 [n=1=A] or Taq.pr8 [n=1=C]). While in theory, an increase in length of the 3'-extension from n=0 to n=1 should decrease the number of amplified fragments four-fold, it is difficult in practice to quantitate the real difference, primarily because of the great degree of general smeariness and poor resolution of the products derived from 5'-anchored simple SSR primers. Clearly, however, the greatest number of bands, along with the highest levels of background, is visible in the lanes representing reactions amplified with Taq.AdF.

Third, thermocycling conditions employing touchdown conditions generally serve to reduce the smearing within the lanes, and generally makes for sharper product bands, in comparison to use of a constant annealing temperature. However, these sharper bands still are accompanied by a great degree of stutter, which hinders precise comparison of polymorphisms between lanes (genomes).

In general, genomic polymorphisms can be detected among individuals from within a species as well as between species; the greater the evolutionary distance between the genomes being compared, the more polymorphisms are expected. Both dominant and codominant polymorphisms are revealed. However, no matter what the specific reaction thermocycling condition, the use of 5'-anchored simple SSR primers in an SSR-to-adaptor amplification is not ideal for identifying new polymorphisms. Even when the annealing conditions are carefully optimized, as in touchdown thermocycling or in a hot start initiation (not shown), the individual co-amplification products resolve on the gels as rather smeary and indistinct. The high degree of stutter apparent on the autoradiographic images counteracts the clarity of the bands attainable using conventional AFLP (Zabeau EP 534,858), and prevents accurate identification of all but the most prominent polymorphisms.

EXAMPLE 2

AMPLIFICATION USING PERFECT COMPOUND SSR PRIMERS IN A MULTIPLEXED SSR-TO-ADAPTOR AMPLIFICATION TO DETECT POLYMORPHISMS AMONG SOYBEAN CULTIVARS

Example 2 illustrates the use of perfect compound SSR-directed primers in an SSR-to-adaptor amplification method similar to that discussed in Example 1 as a means to improve the resolution and increase the level of polymorphism among the multiplexed amplification products. All of the individual compound SSR primers used for these amplifications are listed in Table V.

TABLE V
COMPOUND SSR PRIMERS

Primer Name	Sequence (5'→3')/SEQ ID No.	Length	T _m (50mM Salt)	%GC
<u>Compound SSR-Directed Primers</u>				
(AT) 3.5 (AG) 7.5	TATATATAGAGAGAGAGAGA /63	22	42.3°C	31.80
(TA) 7.5 (GA) 4.5	ATATATATATATATAGAGAGAG /64	24	40.3°C	20.83
(CT) 5 (AT) 7	CTCTCTCTCTATATATATAT /65	24	40.3°C	20.80
(CT) 7.5 (AT) 3.5	TCTCTCTCTCTATATATA /66	22	42.3°C	31.80
(CT) 7.5 (AT) 2	CTCTCTCTCTCTCTATA /67	19	41.6°C	42.00
(AT) 3.5 (GT) 6.5	TATATATGTGTGTGTGTG /68	20	40.5°C	35.00
(AT) 6.5 (GT) 4.5	TATATATATATGTGTGTGTG /69	22	38.5°C	22.70
(AT) 8.5 (GT) 3.5	TATATATATATATGTGTGTG /70	24	38.6°C	16.70
(CA) 4.5 (TA) 7.5	ACACACACATATATATATATAT /71	24	38.4°C	16.70
(CA) 6.5 (TA) 4.5	ACACACACACATATATATAT /72	22	40.4°C	27.27
(CA) 7.5 (TA) 2.5	ACACACACACACATATAT /73	20	40.5°C	35.00
(GT) 7.5 (AT) 2	TGTGTGTGTGTGTATAT /74	19	39.5°C	36.80
(GA) 7.5 (TA) 2	GAGAGAGAGAGAGATAT /75	19	41.6°C	42.10
(TC) 3.5 (AC) 5.5	CTCTCTCACACACACA /76	18	42.9°C	50.00
(TC) 4.5 (AC) 4.5	CTCTCTCTCACACACA /77	18	42.9°C	50.00
(TG) 4.5 (AG) 4.5	GTGTGTGTGAGAGAGAGA /78	18	42.9°C	50.00

5		50.00	
10		42.9°C	
15		42.9°C	
20		42.9°C	
25			
30		18	
35		18	
40		18	
45			
50			
55			
	(CA) 4.5 (GA) 4.5	ACACACACAGAGAGAG /79	
	(TC) 4.5 (TG) 4.5	CTCTCTCTCTGTGTGT /80	
	(GA) 3.5 (GT) 5.5	AGAGAGAGTGTGTGTG /81	

Most of these compound SSR sequences are represented either on sequenced soybean genomic clones that were shown by hybridization to contain one of the dinucleotide repeats that comprise the compound SSR (M. Morgante and C. Andre, unpublished), or on cloned plant and animal sequences that have been entered into public DNA sequence databases (e.g., GenBank; see Table II).

DNA templates were generated essentially as described in Example 1, from the *G. max* cultivars wolverine, NOIR-1, N85-2176, Harrow, CNS, Manchu, Mandarin, Mukden, Richland, Roanoke, Tokyo, PI54-60, and Bonus, as well as from *G. soja* accessions PI81762 and PI440.913. Following double digestion with either Taq I + Pst I or Taq I + Hind III, ligation of double stranded Taq I-Ad and biotin-Pst I-Ad (or biotin-Hind III-Ad) adaptors, and selection of fragments carrying at least one biotinylated Pst I or Hind III end, the amplification reactions were performed using a series of reaction cocktails and a cold start setup, all as described in Example 1. In all cases, only the compound SSR primer was 5'-end labeled using [γ - 33 P]ATP. The amplifications were performed on a Perkin Elmer 9600 thermocycler using a 56°C final annealing temperature touchdown profile:

denature	94°C, 3 min
1 cycle	94°C, 30 sec
	65°C, 30 sec
	72°C, 1 min
11 cycles	94°C, 30 sec
	64.3°C, 30 sec for 1st cycle, then decrease by 0.7°C per cycle for the next 10 cycles to a final 56°C.
	72°C, 1 min
23 cycles	94°C, 30 sec
	56°C, 30 sec
	72°C, 1 min

Following electrophoresis on 6% denaturing polyacrylamide gels in tris-borate-EDTA buffer essentially as described in Example 1, the co-amplified products were visualized by autoradiography after intensifying screen enhanced exposure at -70°C to Kodak Biomax X-ray film.

All of the compound SSR-directed primers listed in Table V have been used in this protocol to detect polymorphisms among different soybean cultivars. All primers used represent perfect compound SSR sequences in which one of the component nucleotides is "in-phase" across the two adjacent dinucleotide repeats. Surprisingly, not all of the primers listed in Table V were equally effective at generating products, and not all generated products with the same degree of polymorphism, even when they were predicted to do so based upon cloned sequence compilations (Table II). The compound primer that produces the greatest number of co-amplified fragments from the soybean genome is (CA)_x(TA)_y (where x and y are multiples of 0.5, but each ≥ 1). Figure 3a shows the amplification products from a (CA)_{7.5}(TA)_{2.5} version of this primer, used in combination with Taq.AdF (containing zero 3'-selective nucleotides) and Taq.pr8 (one 3'-nucleotide, -C). Also shown in Figure 3a are the products from the compound sequence primers (TC)_{4.5}(TG)_{4.5} and (CT)_{7.5}(AT)_{3.5}, which amplify only relatively few fragments even when the Taq I primer is completely nonselective. This result was surprising, since (CT)_x(AT)_y sequences appear to be the second-most abundant class of compound repeat on isolated soybean clones (see Table II). As primers, (CA)_{7.5}(TA)_{2.5} and (CT)_{7.5}(AT)_{3.5} differ primarily in the length of their 3'-(AT)_y repeat. The differing efficiencies of these two primers in otherwise identical amplification reactions may largely be a function of the length of the "leading" 3'-(AT)_y sequence. In contrast, the extremely low number of amplified products resulting from (TC)_{4.5}(TG)_{4.5} is probably the result of low copy number of this compound repeat in the soybean genome (see Table II). Shown in Figure 3b are the products generated using (TG)_{4.5}(AG)_{4.5} and (TC)_{4.5}(AC)_{4.5}, each in combination with the same two Taq I adaptor-specific primers. An intermediate number of products are amplified by each of these two compound SSR sequences.

The amplification products generated with any of the perfect compound SSR-directed primers are a mixture of polymorphic and nonpolymorphic fragments. For example, some of the products from the (CA)_{7.5}(TA)_{2.5} primer are completely nonpolymorphic among all 15 different *G. max* and *G. soja* genotypes tested; however, many of the products reflect either dominant or codominant polymorphisms among these genomes. The products amplified using (CA)_{7.5}(TA)_{2.5} in combination with either Taq.AdF(n=0) or Taq.pr8(n=1=C) from Figure 3a were cataloged:

Band scored as:	AFLP ^a	SSR-to-adaptor amplification (CA) _{7.5} (TA) _{2.5}	
		+ TaqAdF ^b	+ TaqPr8
monomorphic	64	12	14
dominant ^c	34	44	40
codominant ^d	1	8	8
codominant + dominant ^e	3	8	16
TOTAL	102	72	78
% polymorphic products	37%	83%	82%
Expected heterozygosity ^f	0.32	0.44	0.43

^aThe Amplified Fragment Length Polymorphism assay on the same set of 15 soybean cultivars as used for the SSR-to-adaptor amplifications, using two paired adaptor-directed primers corresponding to the Taq I adaptor and Pst I adaptor, respectively (not shown)

^bThe TaqAdF lanes carry a high level of background, and only the most unambiguous polymorphisms were scored

^cA band was scored as polymorphic in the indicated category if at least one of the 15 genotypes showed a difference from the others

^dThese scorings are very conservative, minimum estimates of the true incidence of codominant, allelic products; only the most obvious codominant relationships among the 15 template genotypes were scored. A more accurate measurement of the codominance frequency in these reactions requires analysis of the segregation of potentially allelic bands in a population derived from pairs of these genotypes

^eA band could be scored as both codominant and dominant if it appears to be completely absent from at least one genotype and if it also was represented by at least two size variants in bands that were amplified in other genotypes

^fExpected heterozygosity ($H=1-\sum p_i^2$) for each band (locus) was calculated as the sum of the allele frequencies (p_i) for that locus; an average of H for every polymorphic locus could then be calculated

This set of amplification products does not represent the entirety of the (CA)_{≥7.5}(TA)_{≥2.5} loci in the soybean genome. Within the biotin-selected subset of Taq I + Pst I double digested template fragments, amplification was limited to only those (CA)_{≥7.5}(TA)_{≥2.5} genomic loci for which this SSR is within an amplifiable distance of and is oriented oppositely to a Taq I site, and for which no other Taq I site lies on the other side of the SSR, between the SSR and the "selectable" Pst I site (see Figure 1). The remaining (CA)_{≥7.5}(TA)_{≥2.5} loci in the genome can be amplified from template DNAs constructed using different sets of restriction endonucleases. Figure 4 illustrates that when just one of the two restriction enzymes is changed (Pst I is replaced by Hind III as selectable enzyme site), the pattern of amplified fragments is markedly different from that produced with Pst I + Taq I. Therefore, generating templates restricted with differing combinations of restriction endonucleases, as well as assaying each template preparation with a large set of different compound SSR and adaptor primer combinations, should allow the detection of a greater proportion of the total number of polymorphic SSR loci in any given genome.

In comparison to the low resolution of amplification products generated by SSR-to-adaptor amplification using 5'-anchored simple SSR primers (see Example 1), the alternative use of compound SSR primers allows for a much greater level of product resolution. Each band/product is accompanied by far less stutter, and the overall background in the lanes is noticeably reduced, even when a cold start amplification is used (compare Figure 2 to Figures 3a, 3b). This lower amount of background permits good discrimination of individual products, allowing the assignment of allelic relationships among bands to be made with a greater level of confidence. In some instances, a polymorphism might arise from single nucleotide variation within the region covered by the restriction site or the adaptor-directed primer. Nevertheless, a number of the polymorphisms (most of those categorized as codominant or both codominant and dominant among the 15 cultivars tested) appear to arise from variation in the length of the compound SSR through which primer extension occurs; these are recognizable as codominant polymorphisms whose sizes differ by multiples of the unit length of the repeat. The ability to include repeat length variation as one type of polymorphism that is identifiable by this assay allows for a higher level of polymorphism to be visualized from each amplification reaction. It is possible that this assay will allow for the calculation of genetic distances, both within and between species, with great efficiency. Evidence for this comes from similarity estimates (not shown) made from the data shown in Figures 3a and 3b; distances calculated for pairings of *G. max*-*G. max* genotypes from this data are less than the calculated distances for *G. max*-*G. soja* pairings.

EXAMPLE 3

AMPLIFICATION USING PERFECT COMPOUND SSR PRIMERS IN A MULTIPLEXED SSR-TO-ADAPTOR REACTION FOR GENETIC MAPPING

Example 3 illustrates the use of the SSR-to-adaptor amplification method to generate genetic markers. To deter-

mine whether the polymorphic bands detected between *G. max*, Bonus and *G. soja*, PI 81762 are genetically heritable, and to determine whether these polymorphisms could have utility for genetic mapping, polymorphic products between these strains were scored and mapped to the soybean genome.

5 Genomic template DNAs were made, as described in the MATERIALS AND METHODS, from 66 individuals at the F2 generation from a cross involving the bonus and soja parents (from T. Hymowitz, U. of Illinois). For this example, Taq I and Hind III restriction endonucleases were used for the template DNA double digestions. The digestions, adaptor ligations, and subsequent streptavidin-biotin selections were performed as described in MATERIALS AND METHODS.

10

15

20

25

30

35

40

45

50

55

TABLE VI
F2 Individual Designation

Marker		Parents		F2 Individual Designation																
		Band Name	PI81762	Bonus	1	2	4	6	7	8	9	12	13	14	15	16	17	18	19	20
301		94-23.p4	a	B	a	a	a	a	a	a	a	a	a	a	B	a	a	a	B	a
305		94-23.p1	a	B	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	B
307		94-23.p5	A	b	b	A	b	b	b	b	b	A	A	A	b	b	b	b	A	A
309		94-23.p9	A	b	b	b	b	b	b	b	b	b	b	b	A	A	A	A	A	A
320		94-23.p6	A	b	b	b	b	b	b	b	b	b	b	b	A	A	A	A	A	b
322		94-23.p3	A	B	H	A	B	H	H	B	B	H	H	A	B	m	H	B	A	H
323		94-23.p7	A	b	b	b	b	b	A	b	b	A	b	b	B	A	A	A	b	m
324		94-23.p8a	a	B	a	a	a	B	a	a	a	a	a	a	a	a	B	a	a	a
325		94-23.p8b	a	B	a	a	a	B	a	a	a	a	a	a	a	B	a	a	B	a
326		94-23.p10	a	B	a	a	a	B	b	B	a	a	a	a	a	B	B	a	b	a
329		94-23.p12	A	b	b	b	b	b	b	b	b	b	b	b	b	m	b	b	b	b
331		94-23.p13	A	b	b	A	b	b	A	b	A	b	b	b	b	b	b	b	b	A
333		94-23.p14	A	b	b	b	b	A	b	b	A	b	b	b	b	b	b	b	b	A

Marker		Parents		F2 Individual Designation																
		Band Name	PI81762	Bonus	21	22	25	27	28	29	30	31	33	34	36	37	38	39	40	41
301		94-23.p4	a	B	a	a	B	a	a	B	B	a	B	a	B	a	B	a	a	B
305		94-23.p1	a	B	a	a	b	b	B	a	a	a	a	a	a	a	a	a	B	a
307		94-23.p5	A	b	b	b	b	b	b	A	A	A	b	A	b	b	b	A	b	b
309		94-23.p9	A	b	b	b	b	b	b	b	b	b	A	b	A	A	b	b	A	b
320		94-23.p6	A	b	b	b	b	b	b	b	A	b	b	A	A	b	A	b	b	b
322		94-23.p3	A	B	B	H	H	A	B	H	H	B	H	A	H	B	A	b	B	B
323		94-23.p7	A	b	b	b	m	b	b	b	b	b	b	b	b	b	b	m	b	b
324		94-23.p8a	a	B	a	a	a	a	a	B	a	B	a	a	B	a	B	a	a	a
325		94-23.p8b	a	B	a	a	a	a	a	B	a	a	B	B	a	B	B	a	a	a
326		94-23.p10	a	B	a	a	a	B	a	B	a	a	a	a	a	B	B	a	B	a
329		94-23.p12	A	b	b	b	b	b	b	a	a	b	b	b	a	B	a	b	b	a
331		94-23.p13	A	b	b	b	b	b	A	b	b	b	b	b	A	b	m	b	b	A
333		94-23.p14	A	b	b	b	b	A	b	b	b	b	b	b	A	b	b	A	b	b

5

10

15

20

25

30

35

40

45

50

55

Marker
Name

Band Name	PI81762	Bonus	42	43	44	45	46	47	48	49	50	51	52	53	54	56	57	58
301	94-23.p4	a	B	a	a	a	a	B	a	a	a	a	a	B	B	B	a	B
305	94-23.p1	a	B	a	a	a	B	a	a	a	B	B	B	a	a	a	a	a
307	94-23.p5	A	b	A	b	b	b	A	A	b	b	b	A	b	A	b	b	b
309	94-23.p9	A	b	A	A	A	A	b	b	b	b	A	b	m	A	b	A	b
320	94-23.p6	A	b	b	A	A	b	A	A	B	H	H	A	A	A	m	H	A
322	94-23.p3	A	m	H	A	B	m	A	A	B	B	A	b	A	A	m	b	A
323	94-23.p7	A	b	b	b	b	A	b	b	b	B	A	b	m	b	b	b	b
324	94-23.p8a	a	B	a	a	a	a	a	a	a	B	a	a	a	B	a	a	a
325	94-23.p8b	a	B	a	a	a	a	a	a	a	a	a	a	a	a	B	a	a
326	94-23.p10	a	B	a	a	a	B	a	a	B	a	a	a	a	a	B	a	a
329	94-23.p12	A	b	b	A	A	b	b	A	b	b	A	b	A	b	b	b	b
331	94-23.p13	A	b	b	b	A	b	b	b	b	A	b	A	b	A	b	b	b
333	94-23.p14	A	b	A	b	A	b	b	b	b	b	b	b	A	A	b	A	b

Marker
Name

Band Name	PI81762	Bonus	59	62	64	68	74	76	77	78	80	82	83	84	85	86	91	96
301	94-23.p4	a	B	a	a	a	a	a	a	B	a	a	a	B	B	a	a	B
305	94-23.p1	a	B	a	a	a	B	a	a	B	a	a	a	a	a	a	a	a
307	94-23.p5	A	b	A	b	b	b	b	b	b	b	b	b	b	m	a	b	a
309	94-23.p9	A	b	b	b	b	b	b	A	b	b	b	b	b	m	b	b	b
320	94-23.p6	A	b	b	A	A	H	H	A	A	B	B	A	A	m	b	b	b
322	94-23.p3	A	B	A	A	A	H	H	H	A	B	B	A	A	m	m	m	H
323	94-23.p7	A	b	b	A	A	b	b	b	b	b	b	A	A	m	A	b	m
324	94-23.p8a	a	B	B	B	a	a	a	a	B	a	B	a	a	a	m	a	a
325	94-23.p8b	a	B	a	a	a	a	a	a	B	a	B	a	a	a	a	a	B
326	94-23.p10	a	B	a	a	B	a	a	a	B	B	B	a	a	m	a	a	a
329	94-23.p12	A	b	b	A	b	b	A	b	b	b	b	b	A	m	b	b	b
331	94-23.p13	A	b	b	b	b	A	A	b	b	b	b	A	b	A	b	b	A
333	94-23.p14	A	b	A	b	A	b	b	b	A	b	b	b	A	m	b	b	b

5

10

15

20

25

30

35

40

45

50

55

Marker Name	Band Name	PI81762	Bonus	Linkage Group	Log-Likelihood of Linkage
301	94-23.p4	a	B	LG19	16.9
305	94-23.p1	a	B	LG1	10.8
307	94-23.p5	A	b	LG11	11.4
309	94-23.p9	A	b	LG15	9.4
320	94-23.p6	A	b	LG6	10.4
322	94-23.p3	A	B	LG2	5.4
323	94-23.p7	A	b	LG10	5.8
324	94-23.p8a	a	B	LG14	12.0
325	94-23.p8b	a	B	LG13	11.3
326	94-23.p10	a	B	LG11	14.1
329	94-23.p12	A	b	LG9	3.6
331	94-23.p13	A	b	LG1	3.3
333	94-23.p14	A	b	LG5	11.7

The (CA)_{7.5}(TA)_{2.5} compound SSR-directed primer (Table V) was 5'-end labeled using [γ -³³P]ATP and then used in combination with Taq.pr6 adaptor-directed primer (Table I) for multiplexed amplifications on the parent and F₂ templates. These amplifications utilized the cold start, 56°C touchdown thermocycling protocol detailed in Example 2. Amplification products were resolved on 6% denaturing polyacrylamide gels, then dried and exposed to Kodak Biomax X-ray film. The autoradiograph of these products is shown in Figure 5a. At least 15 amplification products were decisively polymorphic between the parental templates, and demonstrated mendelian segregation among the F₂ individuals. Most of these segregate as dominant polymorphisms, and the probability that each segregated in the F₂ progeny at a 3:1 or a 1:2:1 mendelian ratio, simply by chance alone, was consistently less than 5%. The parental inheritance of each polymorphic product was determined for each of the 66 F₂ individuals, and the scores for 13 of these polymorphisms are shown in Table VI. With this specific primer combination, most of the more unambiguous of the polymorphic bands appear to segregate as dominant markers; only a few polymorphisms appear to segregate codominantly. With other specific primer combinations, however, the incidence of codominant segregation is often higher. Most codominant polymorphisms were more problematic to score in that homozygotes sometimes could not be distinguished from heterozygotes. Therefore, each of the polymorphisms from this amplification were scored with a default assumption of dominance, and instances of true codominance were revealed following mapping.

In order to map these bands to the soybean genome, these inheritance scores were correlated with those of 600 RFLP and single-locus SSR markers previously mapped to the soybean genome (J. A. Rafalski and S. V. Tingey, in: Genetic Maps: Locus Maps of Complex Genomes, 6th Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1993). This standard genetic map (data not shown) was constructed by standard RFLP methodology from analysis of the segregation patterns of many RFLP markers in the same F₂ population as used in this example (for standard RFLP mapping technology, see T. Helentjaris, et al., 1986, *Theor. Appl. Genet.*, 72, 761 which reference is incorporated herein). The basis for genetic mapping analysis is that markers located near to each other in the genome are inherited together in the F₂ progeny, while markers located farther apart are co-inherited less frequently. Segregation analysis and marker map positions were then calculated using a computer segregation analysis program, Map-Maker (E. S. Lander, et al., 1987 *Genomics*, 1, 174) which had been modified by Applicants for the Macintosh. The results indicated that nearly every polymorphic amplification product with an approximate 3:1 dominant segregation ratio, or a 1:2:1 codominant segregation ratio, among the F₂ progeny can be mapped to the soybean genome. This is illustrated in Figure 5b, where 6 polymorphisms from Table VI all map to independent sites on various linkage groups. In total, the 15 polymorphisms mapped from this single primer combination are distributed among 10 different linkage groups; in some instances, two or more polymorphisms localize to linked sites on the same linkage group. The probability that each of the polymorphisms in Figure 5b localize to these positions purely by chance from the observed data varied from 1 in 10^{3.29} to 1 in 10^{16.88}, indicating the strength of these map positions. This example demonstrates that polymorphisms revealed by the present invention, SSR-to-adaptor multiplexed amplification, have utility as genetic markers.

DNA was isolated and templates were made from each of 66 F₂ individuals segregating from a cross between PI81762 and Bonus soybean lines. The genotype of each individual at each of the indicated marker loci was determined as follows: A score of "A" or "B" designates that the locus was inherited from the PI81762 or Bonus parent, respectively. A score of H designates that the locus was inherited from both PI81762 and Bonus. A score of "a" designates that the locus was inherited only from PI81762 or that it was inherited from both Bonus and PI81762. A score of "b" designates that the locus was inherited only from Bonus or that it was inherited from both Bonus and PI81762. A score of "m" indicates missing data.

EXAMPLE 4

AMPLIFICATION USING PERFECT COMPOUND SSR PRIMERS IN A MULTIPLEXED SSR-TO-ADAPTOR REACTION TO DETECT POLYMORPHISMS IN OTHER PLANT AND ANIMAL GENOMES

Example 4 illustrates the use of perfect compound SSR primers for the amplification of genomic DNA from other non-soybean genomes including corn, salmon, human and mouse. Genomic DNA was isolated from the *Z. mays* inbred cultivars, B73, Mo17, ASKC28, LH82, LH119, LH204, AEC272, and CM37, using a urea extraction miniprep procedure (Chen et al., in: The Maize Handbook, M. Freeling and V. Walbot, eds., (1993) pp 526-527, New York). Genomic DNA from five different human sources, as well as from salmon and mouse (BALB/c) were purchased from commercial sources (Sigma, St. Louis, MO or Clontech, Palo Alto CA). All these DNAs were double digested either with Taq I + Pst I or Taq I + Hind III, the restriction fragments ligated to adaptors specific to these restriction sites, and biotin-streptavidin selections performed as described in Examples 1 and 2.

Amplification reactions were performed as described in Example 2, using several individual perfect compound SSR-directed primers (³³P-labeled), each primer in combination with individual Taq I adaptor-directed primers carrying either zero or one 3'-selective nucleotide. The amplifications were performed using cold start, 56°C final touchdown

thermocycling conditions, and the labeled products resolved on 6% denaturing polyacrylamide gels. Examples of the amplification products from these plant and mammalian genomes are shown in Figure 6, panels a, b and c.

In all genomes tested, every specific compound SSR and Taq I adaptor primer combination produced a distinct set of amplification products (a fingerprint). In general, the degree to which any two fingerprints are similar is a function of the evolutionary distance between the individuals. These SSR-to-adaptor amplification product fingerprints have two components that can differ: the absolute number of fragments and their collective pattern on the gel. First, a given SSR-directed primer may generate a completely different relative number of amplification products in one species compared to another species, indicating that the compound SSR locus amplified by that primer may be present in entirely different copy numbers in one phylogenetic group compared to another. For example, the number of co-amplified fragments using the compound repeat, $(CA)_{7.5}(TA)_{2.5}$ appears to be much greater (by at least 2-5 fold) in the mammalian genomes than in soybean or corn (see Figure 6) indicating that mammals contain a greater number of $(CA)_{\geq 7.5}(TA)_{\geq 2.5}$ target loci. This relationship is consistent with the greater estimated frequency of $(CA)_n$ repeats in mammalian versus soybean or corn genomes (Wang et al., *Theor. Applied Genetics*: 88, 1 (1994); Morgante & Olivieri, *Plant J* (1993); Beckmann & Weber *Genomics* 12:627 (1992)). Second, within a narrow phylogenetic group (species or genus), the pattern of amplification products between individuals is generally similar, reflecting the general conservation of restriction sites and SSR loci in the different, yet closely related genomes. Polymorphisms between individuals are detected whenever a particular restriction site or SSR locus that contributes to a given amplified fragment in one genome carries a base substitution, insertion/deletion, or repeat length difference compared to the other genomes of the same species. Virtually no similarities are obvious in the patterns of amplification products between individuals whose evolutionary distance extends beyond the same genus, consistent with the accepted idea that the more diverged the two genomes being compared, the more unlikely they will share common loci. The sets of amplified fragments are entirely different, for example, between individual genomes from human compared to mouse or rat, and the amplification patterns (and likely the set of amplified products) appear to share no similarities between soybean and mammals, or even between soybean and corn.

EXAMPLE 5

EFFECTS OF VARYING THE PRIMER CONSTITUTION AND THERMOCYCLING CONDITIONS TO GENERATE DIFFERENT SETS OF AMPLIFICATION PRODUCTS

Variability in the SSR-to-adaptor amplification reaction, and therefore in the products obtained, results not only from the reaction and thermocycle setup conditions described above, but also from subtleties in the design of the primers and the thermocycling parameters used in these amplifications. Once a particular compound SSR has been chosen as the target locus sequence for this assay, then either partially or entirely different sets of amplification products still can be generated by altering any one of the following primer design criteria: 1) the number and composition of the 3'-extension nucleotide(s) on the adaptor-directed primer; 2) the relative lengths of the two constituent simple repeats that comprise the compound SSR primer; 3) the particular strand of the double-stranded compound SSR locus chosen to correspond to the single-stranded primer (i.e., the directionality of the SSR primer). In addition, the quality of the data generated by a particular amplification is affected by the mode by which thermocycling is initiated.

1. Design of the adaptor-directed primer:

This primer, which corresponds to the synthetic adaptor ligated to the restricted ends of the genomic DNA, can carry a variable number (zero to ten) and arbitrary sequence of nucleotides at its 3'-end. As described by Zabeau (EP 534,858), these variable 3'-nucleotides on the primer anneal specifically to "unknown" sequences that are directly adjacent to the adaptor and restriction site on a genomic DNA fragment, and the recognition by each of only a subset of all possible fragments in the template mixture provides exquisite specificity in the amplification reaction. Since such primers that are otherwise identical in sequence except for differences in the few 3'-most nucleotide(s) can amplify completely nonoverlapping sets of amplification products, they behave much like allele-specific amplification primers (Newton et al., (1989) *Nuc. Acids Res* 17: 2503; Kwok et al., (1990) *Nuc. Acids Res.* 18: 999; Wu et al., (1989) *Proc. Natl. Acad. Sci. USA* 86: 2757), except that their use requires no prior sequence knowledge of the genomic locus to be amplified, and each primer will selectively co-recognize multiple target sites in a template DNA mixture.

In general, the longer the 3'-extension the more selective the primer; as the variable 3'-extension is made longer, the adaptor primer becomes more restricted to recognize a smaller number of potential genomic target sites, leading to a smaller real number of co-amplified products. The addition of each nondegenerate nucleotide onto the 3'-extension leads to approximately 4-fold greater template discrimination. In addition, different single nucleotides at the 3'-most base position(s) give unique template specificities to otherwise identical primers. These principles are illustrated by the examples shown in Figure 7. The ^{33}P -labeled perfect compound SSR primer, $(CA)_{7.5}(TA)_{2.5}$, was used for SSR-

to-adaptor amplification to generate specific products from the genomes of several soybean cultivars, using the experimental conditions described in Example 2. This SSR primer was paired with each of several different Taq I adaptor-directed primers, which all differ only at their 3'-most nucleotide positions (see Table I) :

.AdF	.pr5	.pr6	.pr7	.pr8	.pr9
0	1 (-G)	1 (-A)	1 (-T)	1 (-C)	2 (-AC)

The Taq I adaptor-directed primer with the shortest 3'-extension (zero nucleotides), Taq.AdF, is completely nonselective and generated the largest number of products; the primer with the longest extension (Taq.pr9) is the most selective and resulted in the fewest amplification products. All four of the primers carrying one nucleotide as a 3'-extension amplified approximately the same number of products; however, the set of products from each of these four primers is unique compared to the sets from the other three. The complex pattern of bands amplified using Taq.AdF is actually a composite of all four one-base extension primers. That is, the pattern generated by each of these four primers is approximately a 1/4 subset of the pattern amplified by the zero-base extended primer. Thus, varying both the number and composition of the 3'-selective nucleotides on the adaptor-directed primer is sufficient to generate individual, either partially or completely, nonoverlapping sets of amplification products from the same template when paired with a given SSR-directed primer. The choice of which 3'-extensions to use will depend largely upon relative nucleotide frequencies in a target genome and upon the abundance in the genome of the specific SSR that serves as the other priming site.

2. Relative lengths of the two constituent simple repeats comprising the compound SSR primer:

Every simple and nearly every compound SSR locus in the genome is a double stranded structure whose individual strands carry different permutations of nucleotides. A single-stranded primer that may specifically anneal to one strand at a SSR locus will not anneal to the opposite strand (with the exception of the 9 specific palindromic compound SSR sequences combinations; see Table II). Therefore, a given SSR-directed primer will primer-extend from each genomic target locus in a polar, unidirectional manner, and any compound SSR locus can be recognized and primed from by any of four different primer classes. In addition, each of these four canonical primer classes can include a wide range of individual primers, all differing by the length of the two constituent repeats within the primer. Changes in the lengths of these constituent repeats have profound effects on primer efficacy and the fidelity of reproducible amplifications; in general, the longer the 5'-anchoring repeat relative to the 3'-priming repeat, the better the primer's specificity and priming efficiency in the amplification.

Multiple, individual primers, each differing from the others by the length of its two constituent repeats, have been tested for four different compound SSR sequences: $(CT)_x(AT)_y$, $(AT)_x(AG)_y$, $(CA)_x(TA)_y$, $(AT)_x(GT)_y$ (i.e., the values of x and y are varied for a particular SSR type). Figure 8 shows the results of a test using three different $(CA)_x(TA)_y$ primers, $(CA)_{4.5}(TA)_{7.5}$, $(CA)_{6.5}(TA)_{4.5}$, and $(CA)_{7.5}(TA)_{2.5}$, and three different $(AT)_x(GT)_y$ primers, $(AT)_{3.5}(GT)_{6.5}$, $(AT)_{8.5}(GT)_{2.5}$, and $(AT)_{6.5}(GT)_{4.5}$. All five of these primers were calculated to have a T_m in the range of 38-42°C. Each was 5'-labeled with ^{33}P and paired individually with three different Taq I-directed primers (Taq.AdF, Taq.pr6 and Taq.pr8) in amplification reactions using biotin-selected soybean template DNAs, PI81762 and wolverine, as described in Example 2. Neither of the $(AT)_x(GT)_y$ primers performed very efficiently, although $(AT)_{3.5}(GT)_{6.5}$ generated at least some products, whereas the other two $(AT)_x(GT)_y$ primer versions failed completely to generate any amplification products. In contrast, all three $(CA)_x(TA)_y$ primers were able to generate products, although the number of amplified fragments varied among the three primers. These results demonstrate that the longer the 5'-anchoring repeat and the shorter the 3'-primer extension repeat, the more amplification products are produced. This same conclusion was drawn from similar experiments performed with $(CT)_x(AT)_y$ and $(AT)_x(AG)_y$ primers carrying variable constituent repeat lengths (not shown).

3. Polarity of the single stranded compound SSR-directed primer:

The choice of which strand of a double-stranded compound SSR locus to use as a primer can be extremely critical for determining the success of the SSR-to-adaptor amplification reaction. For some compound SSRs, one strand of the double-stranded SSR was found to serve as an efficient primer whereas the opposite strand failed completely, regardless of the relative lengths of the constituent repeats on the primer. This difference was most extreme for compound SSRs containing a $(AT)_n$ repeat; the only type of $(AT)_n$ containing primer that will lead to efficiently generated amplification products under standard conditions (described in Example 2) is one in which the $(AT)_n$ sequence is very short (1.5-3 repeat units) and is situated as the 3'-primer extension end. Figure 8, for example, illustrates the superior efficiency of $(CA)_x(TA)_y$ primers in contrast to $(AT)_x(GT)_y$ primers, which represent the complementary strands of the same compound SSR. All three of the $(AT)_x(GT)_y$ primers tested were extremely inefficient at generating amplification

products (two were failures), even though the calculated T_m values of all the primers were approximately the same. It is likely that this difference in primer efficiency results from the difference in placement of the $(AT)_n$ stretch within the primer. A and T nucleotides display weak hydrogen bonding during base pairing, and oligonucleotides containing $(AT)_n$ stretches often have self-complementarity artifacts in competition with weak annealing to the template. Therefore, it is likely that an $(AT)_n$ stretch at the 5' end of a primer will serve poorly to anchor the primer to the template, whereas a short $(AT)_n$ at the 3' end will have been well-anchored by the upstream non- $(AT)_n$ portion of the primer.

Primers corresponding to compound SSRs that are devoid of $(AT)_n$ repeats are affected much less by the relative order of the two constituent repeats. For example, the individual primers in the two complementary primer sets, $(TC)_{4.5}(AC)_{4.5}$ versus $(TG)_{4.5}(AG)_{4.5}$ (see Figure 3) and $(CA)_{4.5}(GA)_{4.5}$ vs. $(TC)_{4.5}(AC)_{4.5}$, (not shown) appear to generate amplification products with approximately equivalent efficiencies.

4. Mode of thermocycling initiation

The reaction setup protocol described in the previous examples is essentially a cold start, and allows the possibility for primers to anneal under nonstringent conditions both to template sites that are not necessarily a perfect match, and to multiple, staggered sites within a target locus (the latter leads to a stuttering effect of the amplification products on the gel). This simple reaction setup protocol, nevertheless, routinely was sufficient to generate amplification products that could be distinguished as polymorphic between genomes. In fact, the cold start reaction products from compound SSR directed primers were consistently quite sharp and distinct; however, comparable products derived from 5'-anchored simple SSR primers generally lacked the same clarity and sharpness (compare Figures 2 and 3). Much of this indistinctness and individual product heterogeneity, for both types of SSR-directed primer, could be obviated by the use of a hot start initiation for the thermocycling. A hot start protocol (Chou et al., *Nuc. Acids Res.* 20, 1717, 1992) prevents spurious primer annealing to incorrect template sites prior to the first denaturation, and generates products that resolve more sharply and discretely on the gel.

Otherwise identical amplification reactions were performed using the cold start procedure described in Examples 1 and 2, and also using two different hot start initiation procedures. For one type of hot start, all the components of each amplification reaction were combined as described in MATERIALS AND METHODS, except that the SSR-directed primer (both 5'-end labeled and unlabeled versions) was excluded from the full primer cocktail. The reaction tubes were capped (18.5 μ L reaction volume), and the first denaturation step of the thermocycling was performed (94°C, 3 min). The reactions were then held at 80°C while 1.5 μ L of the appropriate SSR primer (1.0 μ L of 5 ng/ μ L 32 P-labelled plus 0.5 μ L of 50 ng/ μ L unlabelled) was added to each reaction. Exponential amplification was then initiated, using either a constant 58°C annealing temperature or a touchdown (56° or 58°C final) thermocycling protocol. For the second hot start method, Ampliwax PCR-50 gems were used essentially as described by the manufacturer (Stratagene, La Jolla, CA), except that a mixture of 10X buffer, $MgCl_2$, dNTPs, H_2O , and primers (5.4 μ L total) was first heated with the Ampliwax and allowed to cool, then a second mixture (14.6 μ L) consisting of template DNA, AmpliTaq DNA polymerase, PCR buffer and H_2O was added over the top of the re-solidified wax layer (final concentrations of each match those in the standard cold start amplifications). Thermocycling was then initiated, using the touchdown annealing protocol detailed in Example 2.

The amplifications illustrated in Figures 2-8 all employed a cold start protocol. A direct comparison of the two initiation procedures, however, is shown in Figure 9. Two sets of amplifications using the perfect compound SSR primer, $(CA)_{7.5}(TA)_{2.5}$, paired with three different Taq I adaptor primers (Taq.AdF, Taq.pr6, Taq.pr8) were performed using the conditions described in Example 2 (56°C final touchdown temperature); one set was initiated with a standard cold start, the other with the first hot start protocol described above. These results demonstrate that a hot start is superior for generating the most discrete bands with the least amount of stutter, although the cold start is adequate for producing products that nonetheless are discernable as polymorphic between genotypes.

In general, a hot start produced the sharpest product bands on the gel for nearly every SSR-directed primer tested. The most extreme difference between these protocols were observed when 5'-anchored simple SSR primers were used. In fact, a cold start using 5'-anchored simple SSR primers often led to unacceptably smeared and fuzzy product bands. The differences were generally more subtle for compound SSR primers, although some SSR primers (those with long stretches of $(AT)_n$ as the 5'-anchor) failed to produce any product under hot start conditions.

EXAMPLE 6

AMPLIFICATION USING PERFECT COMPOUND SSR-DIRECTED PRIMERS FOR INTER-SSR AMPLIFICATIONS TO DETECT POLYMORPHISMS

Example 6 demonstrates the use of perfect, in-phase compound SSR primers in a single-primer amplification for the production and detection of genetic polymorphisms. Simple SSR sequences containing 3 degenerate bases at the

extreme 5'-end have been shown previously to serve as efficient primers for amplification between neighboring SSR sequences in the genome (Zietkiewicz, et al., *Genomics*, 20, 176, (1994)). Similarly, primers corresponding to compound SSR sequences also are useful and extremely efficient for generating inter-SSR amplification products in a single-primer PCR reaction. Generally, the same compound SSR primers that are most efficient in SSR-to-adaptor amplifications, those with in-phase sequences and corresponding to the most abundant of the compound SSRs in the genome (see Table II) also are the most efficient for generating inter-SSR amplification products.

Figure 10 illustrates the single-primer amplification products obtained using both 5'-anchored simple SSR primers and compound SSR primers, from three different cold start thermocycling protocols: 58°C constant, 58°C touchdown, and 56°C touchdown (described in Examples 1 and 2). SSR primers were 5'-end labeled with [γ -³²P]ATP as described in MATERIALS AND METHODS. Either 20 ng undigested genomic DNA or the standard amount of digested, biotin-selected adaptor modified template DNA were combined with the compound SSR primer (5 ng labeled primer combined with 25 ng unlabeled primer) and the other non-primer components of the amplification reactions described in the previous examples. No adaptor-directed primers were added. These reactions were performed using both hot start (not shown) and cold start conditions although the products resulting from hot start were more discrete and better resolved on the gels.

Figure 10, panel a shows a comparison of the products from undigested genomic DNA for soybean (Bonus and PI 81762) and corn (B73 and CM37) cultivars using the 5'-anchored simple SSR primer, DBD(AC)_{7.5}, generated with the three thermocycling profiles indicated. Panel b illustrates a comparison of the amplification products obtained using the 5'-anchored simple SSR primers, DBD(AC)_{7.5} and HBH(AG)_{8.5}, and perfect compound SSR primers, (AT)_{3.5}(AG)_{7.5} and (AT)_{3.5}(GT)_{6.5}, from both undigested and Taq I + Pst I digested, biotin-selected template DNAs from soybean wolverine and PI 81762. Two different thermocycling methods were used, as indicated.

The undigested DNA templates produced a greater number of amplification products than did the digested templates. The reactions using digested, adaptor-ligated templates served as single-primer controls for the SSR-to-adaptor amplifications described in Examples 1 and 2; relatively few fragments were amplified by the single SSR primer from these cut DNA templates, indicating that most of the amplification products observed in the SSR-to-adaptor reactions are dependent upon the presence of both the SSR sequence and a neighboring adaptor sequence on each digested DNA fragment. The few single-primer products that are visible may result either from bona fide inter-SSR amplification within single DNA fragments, or perhaps from some sort of inter-fragment pairing.

These results demonstrate that compound, as well as simple, SSR primers can generate inter-SSR amplification products from the corn and soybean genomes. The multiple products generated from an individual SSR primer are a mixture of nonpolymorphic and polymorphic fragments. The polymorphic bands between genotypes indicate length differences between or within neighboring SSR sequences in the genome; these fragments are potential markers for genome identification, fingerprint analysis, or marker assisted selection.

EXAMPLE 7

CONVERSION OF SSR-TO-ADAPTOR BAND POLYMORPHISMS TO SINGLE-LOCUS SSR MARKERS

Once a good SSR is found by using the SSR-to-adaptor amplification (or SAMPL) method, it often may be desirable to focus only on this single SSR, or just a few SSRs, for subsequent analysis of a species, and to examine variation at these few SSRs using a more straightforward, nonradioactive, single-locus method.

To accomplish conversion of a band from a SAMPL gel to a single locus marker, this band is first excised from the dried gel (see Figure 11). The DNA from the band is eluted from the gel by heating in 100 μ L H₂O at 95°C, 15 min. The debris is pelleted by centrifugation for 2 min at 12000 rpm, and the DNA in the supernate is precipitated by adding 0.1 volume 3M sodium acetate, pH 5.3, 0.025 volumes 20mg/mL glycogen and 2.5 volumes ethanol, incubating at -70°C 30 min, then centrifuging at 12000rpm, 10 min. The pelleted DNA is washed once in 70% ethanol, air-dried, then resuspended in 10 μ L H₂O. One μ L of this DNA then is used as template for PCR amplification using conditions as described in Example 2, except that the SSR-directed primer and the adaptor-directed primer (each at 1.5ng/ μ L final concentration; corresponding exactly to the primer pair used for the original amplification) each are unlabeled (see Figure 11). These re-amplification products are purified using a Qiagen (Chatsworth, CA) PCR fragment cleanup kit, and the purified DNA fragments either are subcloned into a suitable T-vector (for example, pGEM-T, Promega, Madison, WI) and the insert sequenced using vector-directed primers, or are sequenced directly without subcloning, using the adaptor-directed primer as the sequencing primer. In either case, the DNA sequence for the amplified fragment is obtained, allowing design of the first locus-specific flanking primer (Isfp-1). This primer corresponds to the unique sequence flanking the SSR on the amplified fragment, and is oriented with its 3'-end toward the SSR (Figure 11).

The Isfp-1 primer then is paired with a primer corresponding to the second adaptor used for the initial preparation of the restriction fragment templates, for amplification across the targeted SSR. This amplification can be performed with or without radiolabeling the Isfp-1 primer, although the nonspecific background is reduced with a radiolabeled Isfp-

1 primer. The specifically amplified adaptor-to-lsfp-1 band is excised from the gel and sequenced as described above. From the resulting DNA sequence of the other flanking region of the SSR, a second locus-specific flanking primer, lsfp-2, then is designed. The 3'-end of this primer is oriented toward the SSR, oppositely to lsfp-1.

5 Finally, the lsfp-1 and lsfp-2 unique primers are paired and used for PCR amplification either using restriction fragmented DNA template mixtures or using unrestricted genomic DNA templates. Use of this primer pair generates a locus-specific marker that spans the targeted SSR. However, repeat length variation at this SSR now can be detected quickly and nonradioactively from any undigested genome using these specific flanking region primers.

SEQUENCE LISTING

10

(1) GENERAL INFORMATION:

(i) APPLICANT: MORGANTE, MICHELE
VOGEL, JULIE M.

15

(ii) TITLE OF INVENTION: COMPOUND MICROSATELLITE
PRIMERS FOR THE
DETECTION OF GENETIC
POLYMORPHISMS

20

(iii) NUMBER OF SEQUENCES: 89

(iv) CORRESPONDENCE ADDRESS:

25

(A) ADDRESSEE: E. I. DU PONT DE NEMOURS AND
COMPANY

(B) STREET: 1007 MARKET STREET

(C) CITY: WILMINGTON

(D) STATE: DELAWARE

30

(E) COUNTRY: U.S.A.

(F) ZIP: 19898

(v) COMPUTER READABLE FORM:

35

(A) MEDIUM TYPE: FLOPPY DISK

(B) COMPUTER: IBM PC COMPATIBLE

(C) OPERATING SYSTEM: PC-DOS/MS-DOS

(D) SOFTWARE: PATENT IN RELEASE #1.0,
VERSION 1.25

40

(vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER:

(B) FILING DATE:

45

(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: 08/346,456

50

(B) FILING DATE: 28 NOVEMBER 1994

(viii) ATTORNEY/AGENT INFORMATION:

(A) NAME: FLOYD, LINDA AXAMETHY

55

(B) REGISTRATION NUMBER: 33,692

(C) REFERENCE/DOCKET NUMBER: BB-1064-A

(ix) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: 302-892-8112
(B) TELEFAX: 302-992-7949

(2) INFORMATION FOR SEQ ID NO:1:

5

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 40 base pairs
(B) TYPE: nucleic acid
10 (C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

CACACACACA CACACACACA CATATATATA TATATATATA

40

20

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:

25

(A) LENGTH: 40 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
30 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

35

TATATATATA TATATATATG TGTGTGTGTG TGTGTGTGTG

40

(2) INFORMATION FOR SEQ ID NO:3:

40

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
45 (C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

50 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

AGAGAGAGAG AGAGAGA

17

55 (2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

10

GAGAGAGAGA GAGAGAG

17

(2) INFORMATION FOR SEQ ID NO:5:

15

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

20

(ii) MOLECULE TYPE: DNA (genomic)

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

TCTCTCTCTC TCTCTCT

17

(2) INFORMATION FOR SEQ ID NO:6:

30

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

35

(ii) MOLECULE TYPE: DNA (genomic)

40

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

CTCTCTCTCT CTCTCTC

17

45

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:

50

- (A) LENGTH: 15 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

55

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

ACACACACAC ACACA

15

- 5 (2) INFORMATION FOR SEQ ID NO:8:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 15 base pairs
 (B) TYPE: nucleic acid
 10 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- 15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

CACACACACA CACAC

15

- 20 (2) INFORMATION FOR SEQ ID NO:9:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 15 base pairs
 (B) TYPE: nucleic acid
 25 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- 30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

TGTGTGTGTG TGTGT

15

- (2) INFORMATION FOR SEQ ID NO:10:
- 40 (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 15 base pairs
 (B) TYPE: nucleic acid
 45 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- 50 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GTGTGTGTGT GTGTG

15

- 55 (2) INFORMATION FOR SEQ ID NO:11:
- (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 14 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

10

CCGGTTTTTT TTTT

14

(2) INFORMATION FOR SEQ ID NO:12:

15

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 14 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

20

(ii) MOLECULE TYPE: DNA (genomic)

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GCGCAAAAAA AAAA

14

30

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:

35

- (A) LENGTH: 13 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

40

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

45

ACACACACAC ACA

13

(2) INFORMATION FOR SEQ ID NO:14:

50

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 13 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

55

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

5 **AGAGAGAGAG AGA** 13

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:

- 10 (A) LENGTH: 13 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

20 **TGTGTGTGTG TGT** 13

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:

- 30 (A) LENGTH: 17 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

35 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

40 **CGGCACACAC ACACACA** 17

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:

- 45 (A) LENGTH: 21 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

50 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

55 **CTCGTAGACT GCGTACATGC A** 21

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:

- 5 (A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

10 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

15 **TGTACGCAGT CTAC** 14

(2) INFORMATION FOR SEQ ID NO:19:

20 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
25 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

30 **CTCGTAGACT GCGTACC** 17

35

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:

- 40 (A) LENGTH: 15 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

45 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

50 **AGCTGGTACG CAGTC** 15

(2) INFORMATION FOR SEQ ID NO:21:

55

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 16 base pairs

(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

5 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

10 **GACGATGAGT CCTGAC**

16

(2) INFORMATION FOR SEQ ID NO:22:

15 (i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
20 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

25

CGGTCAGGAC TCAT

14

30 (2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 17 base pairs
35 (B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

40

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

GGAATTCTGG ACTCAGT

45

17

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:

50

(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
55 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

GATCACTGAG TCCAGAATTC C

21

5 (2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:

- 10 (A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

15

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

TGGCCTTTAC AGCGTC

16

20

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:

25

- (A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

30

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

35

TACACGCTGT AAAG

14

(2) INFORMATION FOR SEQ ID NO:27:

40

(i) SEQUENCE CHARACTERISTICS:

- 45 (A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

50

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

CTCGTAGACT GCGTACC

17

55

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:

- 5 (A) LENGTH: 17 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

10 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

15 **CTGCGTACCA GCTTACA**

17

(2) INFORMATION FOR SEQ ID NO:29:

20 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 25 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

30

CTGCGTACCA GCTTACC

17

35 (2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
 40 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

45

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

50 **CTGCGTACCA GCTTAAC**

17

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:

55

- (A) LENGTH: 17 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

CTGCGTACCA GCTTGTC

17

10

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:

15

- (A) LENGTH: 16 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

20

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

CTGCGTACCA GCTTAC

16

25

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:

30

- (A) LENGTH: 16 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

35

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

40

CTGCGTACCA GCTTAA

16

(2) INFORMATION FOR SEQ ID NO:34:

45

(i) SEQUENCE CHARACTERISTICS:

50

- (A) LENGTH: 21 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

55

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

CTCGTAGACT GCGTACATGC A

21

5 (2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:

- 10 (A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

20 GACTGCGTAC ATGCAGAC

18

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:

- 25 (A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

30 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

35

GACTGCGTAC ATGCAGAA

18

(2) INFORMATION FOR SEQ ID NO:37:

40

(i) SEQUENCE CHARACTERISTICS:

- 45 (A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

50 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

GACTGCGTAC ATGCAGCA

18

55

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 18 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

10

GACTGCGTAC ATGCAGTT

18

(2) INFORMATION FOR SEQ ID NO:39:

15

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

20

(ii) MOLECULE TYPE: DNA (genomic)

25 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

GACTGCGTAC ATGCAGA

17

30

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

35

(ii) MOLECULE TYPE: DNA (genomic)

40

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

GACTGCGTAC ATGCAGC

45

17

(2) INFORMATION FOR SEQ ID NO:41:

50 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 16 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

55

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

5 **GACGATGAGT CCTGAC** 16

(2) INFORMATION FOR SEQ ID NO:42:

10 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 15 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- 15 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

20 **ATGAGTCCTG ACCGA** 15

25 (2) INFORMATION FOR SEQ ID NO:43:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
- 30 (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

40 **TGAGTCCTGA CCGAACC** 17

(2) INFORMATION FOR SEQ ID NO:44:

45 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 17 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- 50 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

55

TGAGTCCTGA CCGAACA

17

5

(2) INFORMATION FOR SEQ ID NO:45:

(i) SEQUENCE CHARACTERISTICS:

10

- (A) LENGTH: 17 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

15

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

20

TGAGTCCTGA CCGACAC

17

(2) INFORMATION FOR SEQ ID NO:46:

25

(i) SEQUENCE CHARACTERISTICS:

30

- (A) LENGTH: 17 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:

35

TGAGTCCTGA CCGACAA

17

40

(2) INFORMATION FOR SEQ ID NO:47:

(i) SEQUENCE CHARACTERISTICS:

45

- (A) LENGTH: 16 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

50

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

55

ATGAGTCCTG ACCGAG

16

(2) INFORMATION FOR SEQ ID NO:48:

(i) SEQUENCE CHARACTERISTICS:

- 5 (A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

10 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:

15 **ATGAGTCCTG ACCGAA**

16

(2) INFORMATION FOR SEQ ID NO:49:

20 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
25 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:

30

ATGAGTCCTG ACCGAT

16

35 (2) INFORMATION FOR SEQ ID NO:50:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 16 base pairs
40 (B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

45

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:

ATGAGTCCTG ACCGAC

50

16

(2) INFORMATION FOR SEQ ID NO:51:

(i) SEQUENCE CHARACTERISTICS:

55

- (A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:

TGAGTCCTGA CCGAAC

16

10

(2) INFORMATION FOR SEQ ID NO:52:

(i) SEQUENCE CHARACTERISTICS:

15

(A) LENGTH: 16 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

20

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:

25

TGAGTCCTGA CCGAAA

16

(2) INFORMATION FOR SEQ ID NO:53:

30

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 base pairs

(B) TYPE: nucleic acid

35

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

40

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:

TGAGTCCTGA CCGACA

16

45

(2) INFORMATION FOR SEQ ID NO:54:

(i) SEQUENCE CHARACTERISTICS:

50

(A) LENGTH: 17 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

55

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:

GGAATTCTGG ACTCAGT

17

5 (2) INFORMATION FOR SEQ ID NO:55:

(i) SEQUENCE CHARACTERISTICS:

- 10 (A) LENGTH: 21 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:

GGAATTCTGG ACTCAGTGAT C

20

21

(2) INFORMATION FOR SEQ ID NO:56:

25 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 30 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:

35

TTCTGGACTC AGTGATCT .

18

40 (2) INFORMATION FOR SEQ ID NO:57:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 18 base pairs
 45 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

50

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:

55 TCTGGACTCA GTGATCTT

18

(2) INFORMATION FOR SEQ ID NO:58:

(i) SEQUENCE CHARACTERISTICS:

- 5 (A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

10 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:

15 CTGGACTCAG TGATCTTC

18

(2) INFORMATION FOR SEQ ID NO:59:

20 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 16 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 25 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:

30

TGGCCTTTAC AGCGTC

16

35 (2) INFORMATION FOR SEQ ID NO:60:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 18 base pairs
 40 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

45

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:60:

GCCTTTACAG CGTCTAAT

50

18

(2) INFORMATION FOR SEQ ID NO:61:

(i) SEQUENCE CHARACTERISTICS:

55

- (A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:

CCTTTACAGC GTCTAATC

18

10

(2) INFORMATION FOR SEQ ID NO:62:

(i) SEQUENCE CHARACTERISTICS:

15

- (A) LENGTH: 19 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

20

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:

25

CCTTTACAGC GTCTAATCA

19

(2) INFORMATION FOR SEQ ID NO:63:

(i) SEQUENCE CHARACTERISTICS:

30

- (A) LENGTH: 22 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

35

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:

40

TATATATAGA GAGAGAGAGA GA

22

45

(2) INFORMATION FOR SEQ ID NO:64:

(i) SEQUENCE CHARACTERISTICS:

50

- (A) LENGTH: 24 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

55

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:64:

ATATATATAT ATATAGAGAG AGAG

24

5 (2) INFORMATION FOR SEQ ID NO:65:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 24 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:65:

CTCTCTCTCT ATATATATAT ATAT

24

(2) INFORMATION FOR SEQ ID NO:66:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 22 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:66:

TCTCTCTCTC TCTCTATATA TA

22

(2) INFORMATION FOR SEQ ID NO:67:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 19 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:

CTCTCTCTCT CTCTCTATA

19

(2) INFORMATION FOR SEQ ID NO:68:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:

10

TATATATGTG TGTGTGTGTG

20

15

(2) INFORMATION FOR SEQ ID NO:69:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 22 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

20

(ii) MOLECULE TYPE: DNA (genomic)

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:69:

TATATATATA TATGTGTGTG TG

22

30

(2) INFORMATION FOR SEQ ID NO:70:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 24 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

35

(ii) MOLECULE TYPE: DNA (genomic)

40

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:70:

45

TATATATATA TATATATGTG TGTG

24

50

(2) INFORMATION FOR SEQ ID NO:71:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 24 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

55

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:71:

ACACACACAT ATATATATAT ATAT

24

5

(2) INFORMATION FOR SEQ ID NO:72:

(i) SEQUENCE CHARACTERISTICS:

- 10 (A) LENGTH: 22 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:72:

20

ACACACACAC ACATATATAT AT

22

(2) INFORMATION FOR SEQ ID NO:73:

25

(i) SEQUENCE CHARACTERISTICS:

- 30 (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

35 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:73:

ACACACACAC ACACATATAT

20

40

(2) INFORMATION FOR SEQ ID NO:74:

(i) SEQUENCE CHARACTERISTICS:

- 45 (A) LENGTH: 19 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

50 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:74:

55

TGTGTGTGTG TGTGTATAT

19

(2) INFORMATION FOR SEQ ID NO:75:

(i) SEQUENCE CHARACTERISTICS:

- 5 (A) LENGTH: 19 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

10 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:75:

15 **GAGAGAGAGA GAGAGATAT**

19

(2) INFORMATION FOR SEQ ID NO:76:

20

(i) SEQUENCE CHARACTERISTICS:

- 25 (A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:76:

CTCTCTCACA CACACACA

18

35

(2) INFORMATION FOR SEQ ID NO:77:

(i) SEQUENCE CHARACTERISTICS:

- 40 (A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

45 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:77:

50

CTCTCTCTCA CACACACA

18

(2) INFORMATION FOR SEQ ID NO:78:

55

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 18 base pairs

(B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

5 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:78:

10 GTGTGTGTGA GAGAGAGA

18

(2) INFORMATION FOR SEQ ID NO:79:

15 (i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 20 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:79:

25

ACACACACAG AGAGAGAG

18

30 (2) INFORMATION FOR SEQ ID NO:80:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 18 base pairs
 35 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

40

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:80:

45 CTCTCTCTCT GTGTGTGT

18

(2) INFORMATION FOR SEQ ID NO:81:

50 (i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 18 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 55 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:81:

5 **AGAGAGAGTG TGTGTGTG** 18

(2) INFORMATION FOR SEQ ID NO:82:

10 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- 15 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:82:

TATATATGTG TGTGTGTGTG 20

25 (2) INFORMATION FOR SEQ ID NO:83:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 22 base pairs
- 30 (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

35 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:83:

40 **TATATATATA TATGTGTGTG TG** 22

(2) INFORMATION FOR SEQ ID NO:84:

45 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 22 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- 50 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

55 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:84:

TATATATATA TATATATGTG TG 22

(2) INFORMATION FOR SEQ ID NO:85:

(i) SEQUENCE CHARACTERISTICS:

- 5 (A) LENGTH: 43 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

- 10 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:85:

15 **TATATATATA TATATATATA TATGTGTGTG TGTGTGTGTG TGT** **43**

(2) INFORMATION FOR SEQ ID NO:86:

20

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 43 base pairs
 (B) TYPE: nucleic acid
 25 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

- 30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:86:

TATATATATA TATATATATA TATCACACAC ACACACACAC ACA **43**

35

(2) INFORMATION FOR SEQ ID NO:87:

(i) SEQUENCE CHARACTERISTICS:

- 40 (A) LENGTH: 24 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

- 45 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:87:

50 **TATATATATA TATATACACA CACA** **24**

(2) INFORMATION FOR SEQ ID NO:88:

- 55 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 22 base pairs
 (B) TYPE: nucleic acid

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:88:

TATATATATA CACACACACA CA

22

(2) INFORMATION FOR SEQ ID NO:89:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

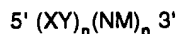
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:89:

TATATACACA CACACACACA

20

Claims

1. An improved method of detecting polymorphisms between two individual nucleic acid samples comprising amplifying segments of nucleic acid from each sample using primer-directed amplification and comparing said amplified segments to detect differences, the improvement comprising wherein at least one of the primers used in said amplification consists of a perfect compound simple sequence repeat in which two different repeating sequences are either directly adjacent or are separated by no more than three intervening bases.
2. The process of Claim 1 wherein said perfect compound simple sequence repeat primer is described by formula I



I

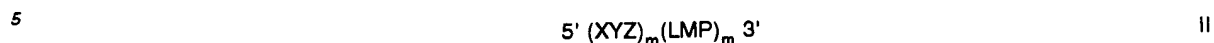
wherein:

n is independently 2-15;
X is A, C, T or G;
Y is A, C, T or G;
N is A, C, T or G;
M is A, C, T or G;

and provided that:

X ≠ Y;
N ≠ M; and
XY ≠ NM.

3. The process of Claim 1 wherein n is independently 4 to 8.
4. The process of Claim 1 wherein said perfect compound simple sequence repeat primer is described by formula II



wherein:

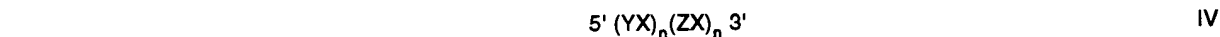
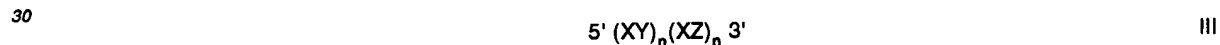
- m is independently 2-10;
 X is A, C, T or G;
 Y is A, C, T or G;
 Z is A, C, T or G;
 M is A, C, T or G;
 N is A, C, T or G;
 P is A, C, T or G;

and provided that: X, Y and Z are not all the same;
 L, M and P are not all the same;

and;

XYZ \neq NMP.

5. The process of Claim 4 wherein m is independently 2 to 4.
6. The process of Claim 1 wherein said perfect compound simple sequence repeat primer is in-phase, described by Formula III or IV



wherein:

- n is independently 2-15;
 X is A, C, T or G;
 Y is A, C, T or G;
 Z is A, C, T or G;

and provided that:

- Y \neq X;
 Z \neq X; and
 Y \neq Z.

7. The process of Claim 6 wherein n is independently 4 to 8.
8. The process of Claim 6 wherein the value of n for the 5' repeating dinucleotide is greater than the value of n for the 3' repeating dinucleotide.
9. The process of Claim 6 wherein said in-phase perfect compound simple sequence repeat primer is selected from the group consisting of:
- $$5' \text{ (AC)}_n \text{ (AT)}_n 3'$$
- $$\text{ (CA)}_n \text{ (TA)}_n$$

(AT)_n(GT)_n
 (TA)_n(TG)_n
 (TA)_n(CA)_n
 (AT)_n(AC)_n
 5 (TG)_n(TA)_n
 (GT)_n(AT)_n
 (TA)_n(GA)_n
 (AT)_n(AG)_n
 (TC)_n(TA)_n
 10 (CT)_n(AT)_n
 (AC)_n(AG)_n
 (CA)_n(GA)_n
 (CT)_n(GT)_n
 (TC)_n(TG)_n
 15 (TG)_n(AG)_n
 (GT)_n(GA)_n
 (CT)_n(CA)_n
 (TC)_n(AC)_n
 (AG)_n(TG)_n
 20 (GA)_n(GT)_n
 (CA)_n(CT)_n
 5' (AC)_n(TC)_n 3'

wherein n is independently 2 to 15.

- 25 10. The process of Claim 1 wherein said primer-directed amplification is performed using a single primer consisting of a perfect compound simple sequence repeat.
11. The process of Claim 1 wherein said perfect compound simple sequence repeat is in-phase.
- 30 12. A process for detecting polymorphisms between two samples of nucleic acid comprising separately treating each nucleic acid sample according to the steps of a-d:
- a) digesting the nucleic acid with at least one restriction enzyme whereby restriction fragments are generated;
- b) ligating adaptor segments to the ends of the restriction fragments of step a);
- 35 c) amplifying the fragments of step b) using primer-directed amplification wherein the amplification primers comprise a first primer consisting of a perfect compound simple sequence repeat as defined in claim 1, and a second primer comprising a sequence which is complementary to an adaptor segment of step b); and
- d) comparing the amplified nucleic acid products of step c) from each nucleic acid sample to detect differences.
- 40 13. The process of Claim 12 in step c) wherein said first primer consists of a perfect compound simple sequence repeat which is in-phase.
14. The process of Claim 12 in step c) wherein said second primer further comprises at the 3' end from 1 to 10 arbitrary nucleotides.
- 45 15. The process of Claim 12 at step a) wherein two different restriction enzymes are used to digest said nucleic acid, one restriction enzyme recognizing a tetranucleotide site on the sample nucleic acid and the other restriction enzyme recognizing a hexanucleotide site on the sample nucleic acid; and further wherein at step b) two different adaptor segments are ligated to the restriction fragments generated at step a).
- 50 16. The process of Claim 15 wherein at step b) one of the two adaptor segments carries a member of a binding pair.
17. The process of Claim 16 wherein said member of a binding pair is biotin.
- 55 18. The process of Claim 16 further comprising an additional step performed after step b):
- b) (i) separating those fragments of step b) which carry a member of a binding pair from those fragments of step b) which do not carry a member of a binding pair; and further at step c) wherein only those fragments at

step b) (i) which carry a member of a binding pair are amplified according to step c).

19. The process of Claim 12 at step c) wherein said first primer carries a reporter molecule.
- 5 20. The process of Claim 19 wherein said reporter is ^{32}P or ^{33}P .
21. The process of Claim 12 at step c) wherein said amplification is performed using a touchdown thermocycling protocol.
- 10 22. The process of Claim 12 at step c) wherein said amplification is initiated using a hot start protocol.
23. The process of Claim 13 wherein said in-phase perfect compound simple sequence repeat is selected from the group consisting of:
- 15 5' (AC)_n(AT)_n 3'
- (CA)_n(TA)_n
- (AT)_n(GT)_n
- (TA)_n(TG)_n
- (TA)_n(CA)_n
- (AT)_n(AC)_n
- 20 (TG)_n(TA)_n
- (GT)_n(AT)_n
- (TA)_n(GA)_n
- (AT)_n(AG)_n
- (TC)_n(TA)_n
- 25 (CT)_n(AT)_n
- (AC)_n(AG)_n
- (CA)_n(GA)_n
- (CT)_n(GT)_n
- (TC)_n(TG)_n
- 30 (TG)_n(AG)_n
- (GT)_n(GA)_n
- (CT)_n(CA)_n
- (CA)_n(CT)_n
- (AG)_n(TG)_n
- 35 (GA)_n(CT)_n
- (CA)_n(CT)_n
- 5' (AC)_n(TC)_n 3'
- wherein n is independently 2 to 15.
- 40 24. The process of Claim 23 wherein the value of n for the 5' repeating dinucleotide is greater than the value of n for the 3' repeating dinucleotide.
25. A process for detecting polymorphisms between two samples of nucleic acid comprising separately treating each nucleic acid sample according to the steps of a-d:
- 45 a) digesting the nucleic acid with at least one restriction enzyme whereby restriction fragments are generated;
- b) ligating adaptor segments to the ends of the restriction fragments of step a);
- c) amplifying the fragments of step b) using primer-directed amplification wherein the amplification primers comprise a first primer consisting of a simple sequence repeating region at the 3' end and a degenerate nucleotide region at the 5' end; and a second primer comprising a sequence which is complementary to an adaptor segment of step b); and
- 50 d) comparing the amplified nucleic acid products of step c) from each nucleic acid sample to detect differences.
26. The process of Claim 25 at step c) wherein said first primer is described by Formula V,
- 55



V

wherein:

X is A, C, T or G;

Y is A, C, T or G;

X \neq Y;

r is 2 to 6; and

n is 2 to 15.

27. The process of Claim 25 at step c) wherein said first primer is described by Formula VI;



VI

wherein:

X is A, C, T or G;

Y is A, C, T or G;

Z is A, C, T or G;

X, Y, and Z are not all the same;

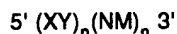
r is 2 to 6; and

m is 2 to 10.

Patentansprüche

1. Verbessertes Verfahren zum Nachweisen von Polymorphismen zwischen zwei individuellen Nucleinsäureproben, umfassend Amplifizieren von Nucleinsäuresegmenten aus jeder Probe unter Verwenden von Primer gerichteter Amplifikation und Vergleichen der amplifizierten Segmente unter Nachweisen von Unterschieden, wobei die Verbesserung umfaßt, wobei mindestens einer der bei der Amplifikation verwendeten Primer aus einer Perfekt-Verbindungs-Einfach-Sequenz-Wiederholung besteht, in der zwei unterschiedliche Wiederholungs-Sequenzen durch nicht mehr als drei dazwischenliegende Basen getrennt sind.

2. Verfahren nach Anspruch 1, wobei der Perfekt-Verbindungs-Einfach-Sequenz-Wiederholungs-Primer durch die Formel 1 beschrieben wird,



I

wobei

n unabhängig 2-15 ist;

X A, C, T oder G ist;

Y A, C, T oder G ist;

N A, C, T oder G ist;

M A, C, T oder G ist;

und unter der Voraussetzung daß:

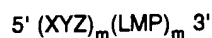
X \neq Y;

N \neq M; und

XY \neq NM.

3. Verfahren nach Anspruch 1, wobei n unabhängig 4 bis 8 ist.

4. Verfahren nach Anspruch 1, wobei der Perfekt-Verbindungs-Einfach-Sequenz-Wiederholungs-Primer durch die Formel II



II

5 beschrieben wird,
wobei

10 m unabhängig 2-10 ist;
X A, C, T oder G ist;
Y A, C, T oder G ist;
Z A, C, T oder G ist;
M A, C, T oder G ist;
N A, C, T oder G ist;
P A, C, T oder G ist;

15 und unter der Voraussetzung, daß

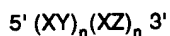
X, Y und Z nicht alle gleich sind;
L, M und P nicht alle gleich sind; und
XYZ \neq NMP.

20

5. Verfahren nach Anspruch 4, wobei m unabhängig 2 bis 4 ist.

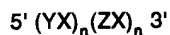
6. Verfahren nach Anspruch 1, wobei der Perfekt-Verbindungs-Einfach-Sequenz-Wiederholungs-Primer In-Phase ist, beschrieben durch die Formel III oder IV

25



III

30



IV

wobei

35 n unabhängig 2-15 ist;
X A, C, T oder G ist;
Y A, C, T oder G ist;
Z A, C, T oder G ist;

40 und unter der Voraussetzung, daß

Y \neq X;
Z \neq X; und
Y \neq Z.

45

7. Verfahren nach Anspruch 6, wobei n unabhängig 4 bis 8 ist.

8. Verfahren nach Anspruch 6, wobei der Wert von n für das 5' Wiederholungsdinucleotid größer als der Wert von n für das 3' Wiederholungsdinucleotid ist.

50

9. Verfahren nach Anspruch 6, wobei der In-Phasen-Perfekt-Verbindungs-Einfach-Sequenz-Wiederholung-Primer ausgewählt ist aus der Gruppe aus:

55 5' (AC)_n(AT)_n 3'
(CA)_n(TA)_n
(AT)_n(GT)_n
(TA)_n(TG)_n
(TA)_n(CA)_n
(AT)_n(AC)_n

5
 10
 15
 20
 25
 30
 35
 40
 45
 50
 55

$(TG)_n(TA)_n$
 $(GT)_n(AT)_n$
 $(TA)_n(GA)_n$
 $(AT)_n(AG)_n$
 $(TC)_n(TA)_n$
 $(CT)_n(AT)_n$
 $(AC)_n(AG)_n$
 $(CA)_n(GA)_n$
 $(CT)_n(GT)_n$
 $(TC)_n(TG)_n$
 $(TG)_n(AG)_n$
 $(GT)_n(GA)_n$
 $(CT)_n(CA)_n$
 $(TC)_n(AC)_n$
 $(AG)_n(TG)_n$
 $(GA)_n(GT)_n$
 $(CA)_n(CT)_n$
 $5' (AC)_n(TC)_n 3'$

wobei n unabhängig 2 bis 15 ist.

10. Verfahren nach Anspruch 1, wobei die Primer gerichtete Amplifikation unter Verwenden eines Einzelprimers bestehend aus einer Perfekt-Verbindungs-Einfach-Sequenz-Wiederholung durchgeführt wird.
11. Verfahren nach Anspruch 1, wobei die Perfekt-Verbindungs-Einfach-Sequenz-Wiederholung In-Phase ist.
12. Verfahren zum Nachweisen von Polymorphismen zwischen zwei Proben von Nucleinsäure, umfassend separates Behandeln jeder Nucleinsäureprobe gemäß den Stufen a-d:
 - a) Verdauen der Nucleinsäure mit mindestens einem Restriktionsenzym, wodurch Restriktionsfragmente erzeugt werden;
 - b) Ligasieren von Adaptorsegmenten an die Enden der Restriktionsfragmente der Stufe a);
 - c) Amplifizieren der Fragmente der Stufe b) unter Verwenden von Primer gerichteter Amplifikation, wobei die Amplifikationsprimer einen ersten Primer umfassend bestehend aus einer Perfekt-Verbindungs-Einfach-Sequenz-Wiederholung, nach Anspruch 1, und einen zweiten Primer, umfassend eine Sequenz, welche komplementär zu einem Adaptorsegment der Stufe b) ist; und
 - d) Vergleichen der amplifizierten Nucleinsäureprodukte der Stufe c) von jeder Nucleinsäureprobe unter Nachweisen von Unterschieden.
13. Verfahren nach Anspruch 12 in Stufe c), wobei der erste Primer aus einer Perfekt-Verbindungs-Einfach-Sequenz-Wiederholung besteht, die In-Phase ist.
14. Verfahren nach Anspruch 12 in Stufe c), wobei der zweite Primer ferner an dem 3' Ende 1 bis 10 willkürliche Nucleotide umfaßt.
15. Verfahren nach Anspruch 12 bei Stufe a), wobei zwei unterschiedliche Restriktionsenzyme verwendet werden um die Nucleinsäure zu verdauen, ein Restriktionsenzym eine Tetranucleotidstelle auf der Probennucleinsäure erkennt und das andere Restriktionsenzym eine Hexanucleotidstelle auf der Probennucleinsäure erkennt; und wobei ferner bei Stufe b) zwei unterschiedliche Adaptorsegmente an die bei Stufe a) erzeugten Restriktionsfragmente ligasiert werden.
16. Verfahren nach Anspruch 15, wobei bei Stufe b) eines der zwei Adaptorsegmente ein Glied eines Bindungspaares trägt.
17. Verfahren nach Anspruch 16, wobei das Glied eines Bindungspaares Biotin ist.
18. Verfahren nach Anspruch 16, welches ferner eine nach Stufe b) durchgeführte zusätzliche Stufe umfaßt:
 - b)(i) Trennen jener Fragmente der Stufe b), welche ein Glied eines Bindungspaares tragen, von denjenigen

Fragmenten der Stufe b), welche nicht ein Glied eines Bindungspaares tragen; und ferner bei Stufe c), wo nur diejenigen Fragmente bei Stufe b)(i) welche ein Glied eines Bindungspaares tragen, gemäß Stufe c) amplifiziert werden.

- 5 19. Verfahren nach Anspruch 12 bei Stufe c), wobei der erste Primer ein Reportermolekül trägt.
20. Verfahren nach Anspruch 19, wobei der Reporter ^{32}P oder ^{33}P ist.
- 10 21. Verfahren nach Anspruch 12 bei Stufe c), wobei die Amplifikation unter Verwenden eines Touchdown Thermocycling Protokolls durchgeführt wird.
22. Verfahren nach Anspruch 12 bei Stufe c), wobei die Amplifikation unter Verwenden eines Heiß-Start Protokolls initiiert wird.
- 15 23. Verfahren nach Anspruch 13, wobei die In-Phase-Perfekt-Verbindungs-Einfach-Sequenz-Wiederholung ausgewählt ist aus der Gruppe aus:
- 5'(AC)_n(AT)_n 3'
- (CA)_n(TA)_n
- (AT)_n(GT)_n
- 20 (TA)_n(TG)_n
- (TA)_n(CA)_n
- (AT)_n(AC)_n
- (TG)_n(TA)_n
- (GT)_n(AT)_n
- 25 (TA)_n(GA)_n
- (AT)_n(AG)_n
- (TC)_n(TA)_n
- (CT)_n(AT)_n
- (AC)_n(AG)_n
- 30 (CA)_n(GA)_n
- (CT)_n(GT)_n
- (TC)_n(TG)_n
- (TG)_n(AG)_n
- (GT)_n(GA)_n
- 35 (CT)_n(CA)_n
- (CA)_n(CT)_n
- (AG)_n(TG)_n
- (GA)_n(CT)_n
- (CA)_n(CT)_n
- 40 5' (AC)_n(TC)_n 3'
- wobei n unabhängig 2 bis 15 ist.
24. Verfahren nach Anspruch 23, wobei der Wert von n für das 5' Wiederholungsdinucleotid größer als der Wert von n für das 3' Wiederholungsdinucleotid ist.
- 45 25. Verfahren zum Nachweisen von Polymorphismen zwischen zwei Proben von Nucleinsäure, umfassend separates Behandeln jeder Nucleinsäurenprobe gemäß den Stufen a-d:
- a) Verdauen der Nucleinsäure mit mindestens einem Restriktionsenzym, wodurch Restriktionsfragmente erzeugt werden;
- 50 b) Ligasieren von Adaptorsegmenten an die Enden der Restriktionsfragmente der Stufe a);
- c) Amplifizieren der Fragmente der Stufe b) unter Verwenden von Primer gerichteter Amplifikation, wobei die Amplifikationsprimer einen ersten Primer, bestehend aus einer Einfach-Sequenz-Wiederholungs-Region an dem 3' Ende und einer degenerierten Nucleotidregion an dem 5' Ende; und einen zweiten Primer, umfassend
- 55 eine Sequenz, welche komplementär zu einem Adaptorsegment der Stufe b) ist, umfassen; und
- d) Vergleichen der amplifizierten Nucleinsäureprodukte der Stufe c) von jeder Nucleinsäurenprobe zum Nachweisen von Unterschieden.

26. Verfahren nach Anspruch 25 bei Stufe c), wobei der erste Primer durch Formel V beschrieben wird:



V

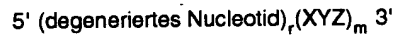
5

wobei:

- 10
X A, C, T oder G ist;
Y A, C, T oder G ist;
X \neq Y;
r 2 bis 6 ist; und
n 2 bis 15 ist.

27. Verfahren nach Anspruch 25 bei Stufe c), wobei der erste Primer durch Formel VI beschrieben wird,

15



VI

20

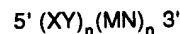
wobei:

- 25
X A, C, T oder G ist;
Y A, C, T oder G ist;
Z A, C, T oder G ist;
X, Y und Z nicht alle gleich sind;
r 2 bis 6 ist; und
m 2 bis 10 ist.

Revendications

30

1. Un procédé amélioré de détection de polymorphismes entre deux échantillons individuels d'acides nucléiques comprenant l'amplification des segments d'acide nucléique de chaque échantillon en utilisant une amplification orientée par une amorce et en comparant lesdits segments amplifiés pour détecter les différences, l'amélioration comprenant qu'au moins une des amorces utilisées dans ladite amplification consiste en une répétition parfaite de séquence simple composite dans laquelle deux séquences répétitives ne sont pas séparées par plus de trois bases intermédiaires.
- 35
2. Le procédé selon la revendication 1, dans lequel ladite amorce répétitive de séquence simple composite parfaite est décrite par la formule I :
- 40



I

45

dans laquelle :

- n vaut indépendamment de 2 à 15 ;
X représente A, C, T ou G ;
Y représente A, C, T ou G ;
50 N représente A, C, T ou G ;
M représente A, C, T ou G ;

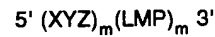
pourvu que :

- 55 X \neq Y ;
N \neq M ; et
XY \neq NM.

3. Le procédé selon la revendication 1, dans lequel n vaut indépendamment de 4 à 8.

4. Le procédé selon la revendication 1, dans lequel ladite amorce répétitive à séquence simple composite parfaite est décrite par la formule II :

5



II

dans laquelle :

10

m vaut indépendamment de 2 à 10 ;

X représente A, C, T ou G ;

Y représente A, C, T ou G ;

Z représente A, C, T ou G ;

15

M représente A, C, T ou G ;

N représente A, C, T ou G ;

P représente A, C, T ou G ;

pourvu que :

20

X, Y et Z ne soient pas tous les mêmes ;

L, M et P ne soient pas tous les mêmes ;

et

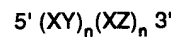
25

XYZ ≠ NMP.

5. Le procédé selon la revendication 4, dans lequel m vaut indépendamment de 2 à 4.

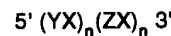
6. Le procédé selon la revendication 1, dans lequel ladite amorce répétitive à séquence simple composite parfaite est en phase, décrite par la formule III ou IV :

30



III

35



IV

dans lesquelles :

40

n vaut indépendamment de 2 à 15 ;

X représente A, C, T ou G ;

Y représente A, C, T ou G ;

Z représente A, C, T ou G ;

45

pourvu que :

X ≠ Y ;

Z ≠ X ; et

Y ≠ Z.

50

7. Le procédé selon la revendication 6, dans lequel n vaut indépendamment de 4 à 8.

8. Le procédé selon la revendication 6, dans lequel la valeur de n pour les dinucléotides répétitifs 5' est supérieure à la valeur de n pour le dinucléotide répétitif 3'.

55

9. Le procédé selon la revendication 6, dans lequel ladite amorce répétitive à séquence simple composite parfaite

en phase est choisie dans le groupe consistant en :

- 5' (AC)_n(AT)_n 3'
- (CA)_n(TA)_n
- (AT)_n(GT)_n
- 5 (TA)_n(TG)_n
- (TA)_n(CA)_n
- (AT)_n(AC)_n
- (TG)_n(TA)_n
- (GT)_n(AT)_n
- 10 (TA)_n(GA)_n
- (AT)_n(AG)_n
- (TC)_n(TA)_n
- (CT)_n(AT)_n
- (AC)_n(AG)_n
- 15 (CA)_n(GA)_n
- (CT)_n(GT)_n
- (TC)_n(TG)_n
- (TG)_n(AG)_n
- (GT)_n(GA)_n
- 20 (CT)_n(CA)_n
- (TC)_n(AC)_n
- (AG)_n(TG)_n
- (GA)_n(GT)_n
- (CA)_n(CT)_n
- 25 5' (AC)_n(TC)_n 3'

dans lequel n est indépendamment égal à 2 à 15.

10. Le procédé selon la revendication 1 dans lequel ladite amplification orientée par amorce est mise en oeuvre en utilisant une amorce unique consistant en un motif répétitif à séquence simple composite parfait.
- 30 11. Le procédé selon la revendication 1, dans lequel ledit motif répétitif à séquence simple composite parfait est en phase.
12. Un procédé de détection de polymorphismes entre deux échantillons d'acide nucléique comprenant le traitement séparément de chaque échantillon d'acide nucléique selon les étapes a à d :
- 35 a) digestion de l'acide nucléique en présence d'au moins une enzyme de restriction, ce qui fait apparaître des fragments de restriction ;
- b) ligation de segments adaptateurs aux extrémités des fragments de restriction de l'étape a) ;
- 40 c) amplification des fragments de l'étape b) en utilisant une amplification orientée par amorce dans laquelle les amorces d'amplification comprennent une première amorce consistant en un motif répétitif à séquence simple composite parfait telle que définie dans la revendication 1, et une seconde amorce comprenant une séquence qui est complémentaire d'un segment adaptateur de l'étape b) ; et
- d) comparaison des produits d'acide nucléique amplifié de l'étape c) provenant de chaque échantillon d'acide nucléique pour en détecter les différences.
- 45 13. Le procédé selon la revendication 12 dans l'étape c), dans lequel ladite première amorce est constituée d'un motif répétitif en séquence simple composite parfait qui est en phase.
- 50 14. Le procédé selon la revendication 12 dans l'étape c), dans lequel ladite seconde amorce comprend en outre, à l'extrémité 3', de 1 à 10 nucléotides arbitraires.
15. Le procédé selon la revendication 12 dans l'étape a), dans lequel deux enzymes de restriction différentes sont utilisées pour digérer ledit acide nucléique. une enzyme de restriction reconnaissant un site tétranucléotidique sur l'échantillon d'acide nucléique et l'autre enzyme de restriction reconnaissant un site hexanucléotidique sur l'échantillon d'acide nucléique ; et ensuite, dans lequel, à l'étape b), deux segments d'adaptateur différents font l'objet d'une ligation sur des fragments de restriction engendrés à l'étape a).
- 55

16. Le procédé selon la revendication 15, dans lequel, à l'étape b), l'un des deux segments adaptateurs porte un élément d'une paire de liaisons.
17. Le procédé selon la revendication 16, dans lequel ledit élément d'une paire de liaisons est la biotine.
18. Le procédé selon la revendication 16, comprenant en outre une étape additionnelle mise en oeuvre après l'étape b) :
- b) (i) séparation des fragments de l'étape b) qui portent un élément d'une paire de liaisons provenant des fragments de l'étape b) qui ne portent pas un élément d'une paire de liaisons ; et
- en outre, à l'étape c), dans laquelle seuls les fragments de l'étape b) (i) qui portent un élément de la paire de liaison sont amplifiés conformément à l'étape (c).
19. Le procédé de la revendication 12, dans l'étape c), dans lequel ladite première amorce porte une molécule marqueur.
20. Le procédé de la revendication 19, dans lequel ledit marqueur est ^{32}P ou ^{33}P .
21. Le procédé de la revendication 12, dans l'étape c), dans lequel ladite amplification est mise en oeuvre en utilisant un protocole de cycle thermique avec point bas.
22. Le procédé de la revendication 12, dans l'étape c), dans lequel ladite amplification est initiée en utilisant un protocole avec démarrage à chaud.
23. Le procédé de la revendication 13, dans lequel ledit motif répétitif de la séquence simple composite parfait en phase est choisie dans le groupe consistant en :
- 5' (AC)_n(AT)_n 3'
 (CA)_n(TA)_n
 (AT)_n(GT)_n
 (TA)_n(TG)_n
 (TA)_n(CA)_n
 (AT)_n(AC)_n
 (TG)_n(TA)_n
 (GT)_n(AT)_n
 (TA)_n(GA)_n
 (AT)_n(AG)_n
 (TC)_n(TA)_n
 (CT)_n(AT)_n
 (AC)_n(AG)_n
 (CA)_n(GA)_n
 (CT)_n(GT)_n
 (TC)_n(TG)_n
 (TG)_n(AG)_n
 (GT)_n(GA)_n
 (CT)_n(CA)_n
 (CA)_n(CT)_n
 (AG)_n(TG)_n
 (GA)_n(CT)_n
 (CA)_n(CT)_n
 5' (AC)_n(TC)_n 3'
- dans lequel n est indépendamment égal à 2 à 15.
24. Le procédé selon la revendication 23, dans lequel la valeur de n pour ledit nucléotide répétitif en 5' est supérieure à la valeur de n pour ledit nucléotide répétitif en 3'.
25. Un procédé de détection de polymorphismes entre deux échantillons d'acide nucléique comprenant traitement séparément de chaque échantillon d'acide nucléique conformément aux étapes a à d :

- a) digestion de l'acide nucléique en présence d'au moins une enzyme de restriction, ce qui fait apparaître des fragments de restriction ;
 b) ligation de segments adaptateurs aux extrémités des fragments de restriction de l'étape a) ;
 c) amplification des fragments de l'étape b) en utilisant une amplification orientée par amorce dans laquelle les amorces d'amplification comprennent une première amorce consistant en une région répétitive en séquence simple à l'extrémité 3' et une région de nucléotide dégénérée à l'extrémité 5'; et une seconde amorce comprenant une séquence qui est complémentaire d'un segment adaptateur de l'étape b) ; et
 d) comparaison des produits d'acide nucléique amplifié de l'étape c) provenant de chaque échantillon d'acide nucléique pour en détecter les différences.

26. Le procédé selon la revendication 25 à l'étape c), dans lequel ladite première amorce est décrite par la formule V :



dans laquelle :

X représente A, C, T ou G ;

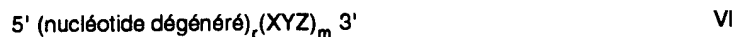
Y représente A, C, T ou G ;

$X \neq Y$

r vaut de 2 à 6 ; et

n vaut de 2 à 15.

27. Le procédé selon la revendication 25 à l'étape c), dans lequel ladite première amorce est décrite par la formule VI :



dans laquelle :

X représente A, C, T ou G ;

Y représente A, C, T ou G ;

Z représente A, C, T ou G ;

X, Y et Z ne sont pas tous identiques ;

r vaut de 2 à 6 ; et

m vaut de 2 à 10.

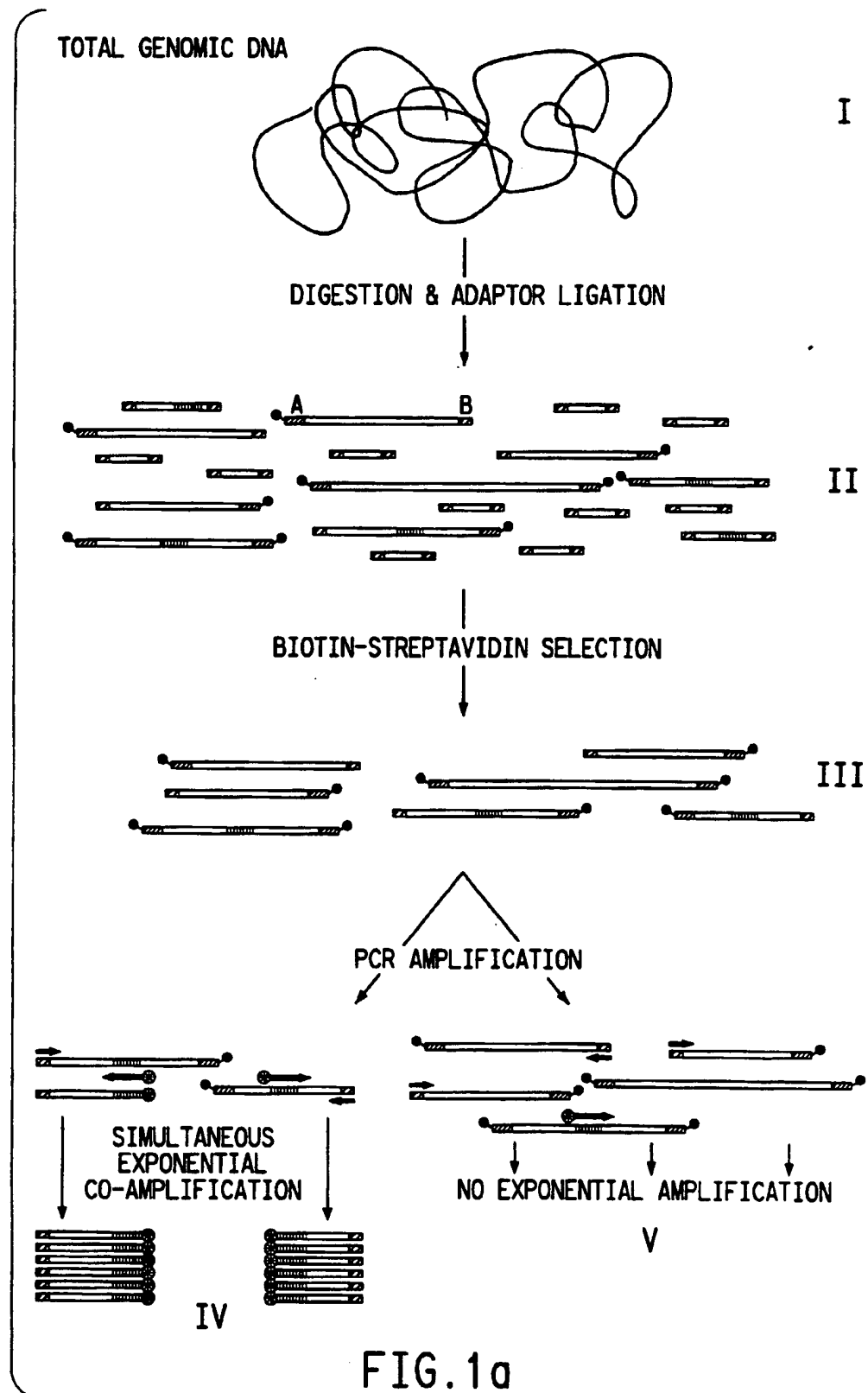


FIG. 1b

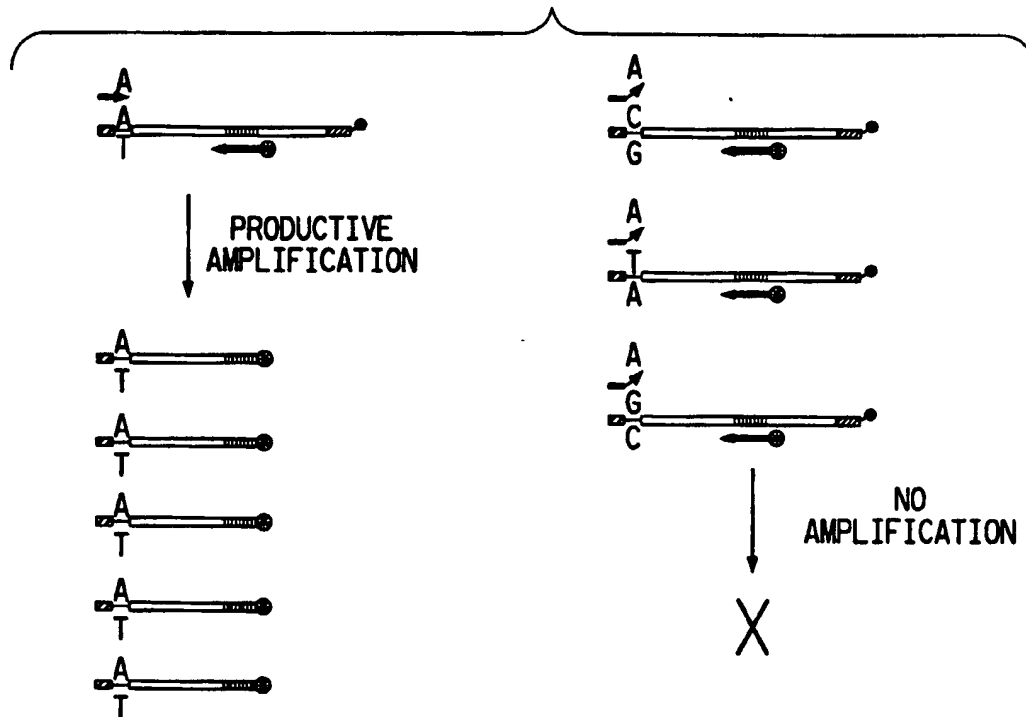


FIG. 1c

IN-PHASE PERFECT COMPOUND SSR LOCUS: (AT)_{11.5}(GT)₁₀

PRIMER:

(AT)_{2.5}(GT)_{6.5}

(AT)_{6.5}(GT)_{4.5}

(AT)_{8.5}(GT)_{3.5}

TATATATGTGTGTGTGTGTG →
 TATATATATATATGTGTGTGTG →
 TATATATATATATATGTGTG →
 5'..NNTATATATATATATATATATATATGTGTGTGTGTGTGTGTGTNN.. 3'
 3'..NNTATATATATATATATATATATATCACACACACACACACACANN.. 5'

(CA)_{4.5}(TA)_{7.5}

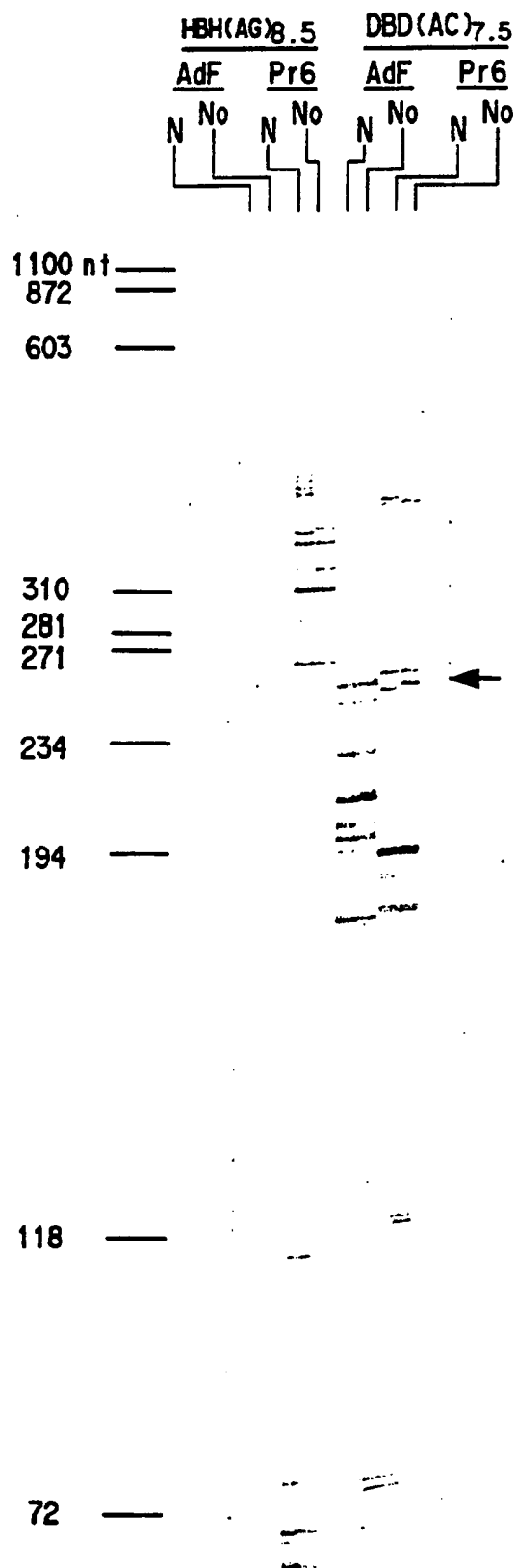
(CA)_{6.5}(TA)_{4.5}

(CA)_{4.5}(TA)_{7.5}

← TATATATATATATATACACACACA

← TATATATATACACACACACACA

← TATATACACACACACACACA



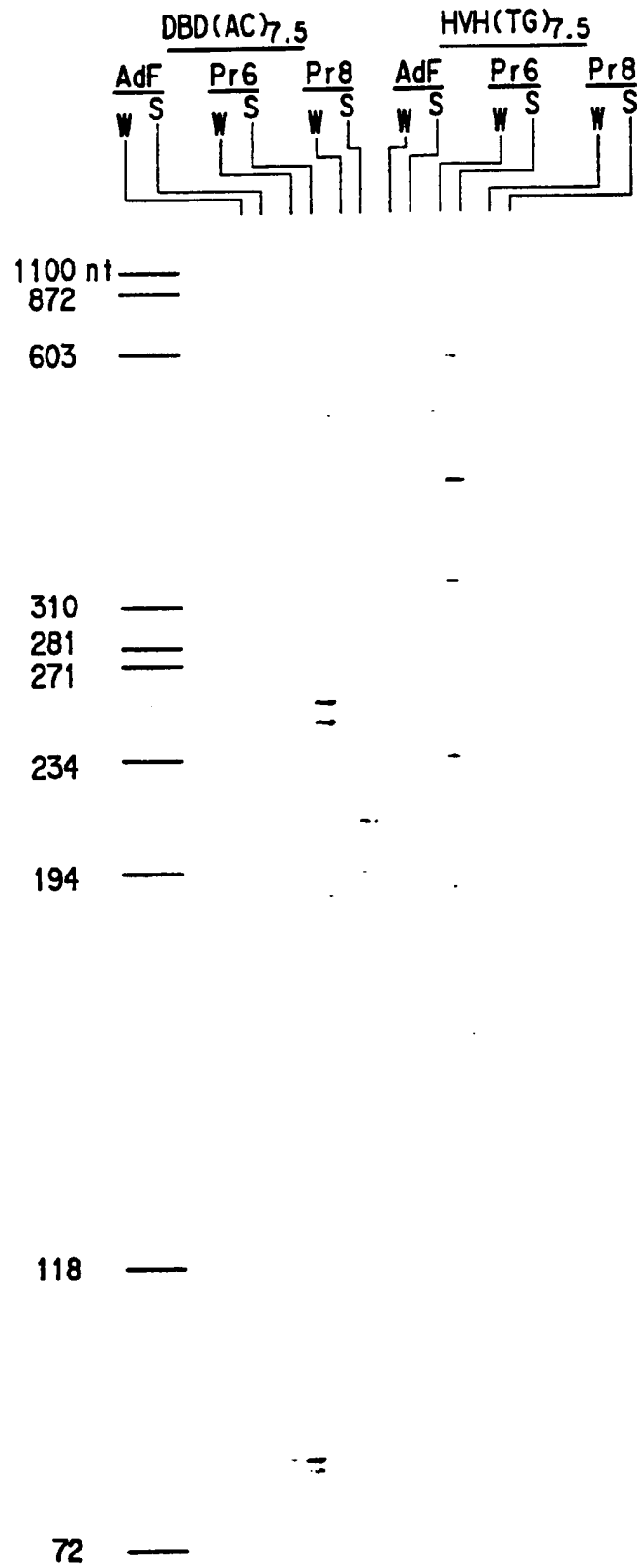


FIG.2b

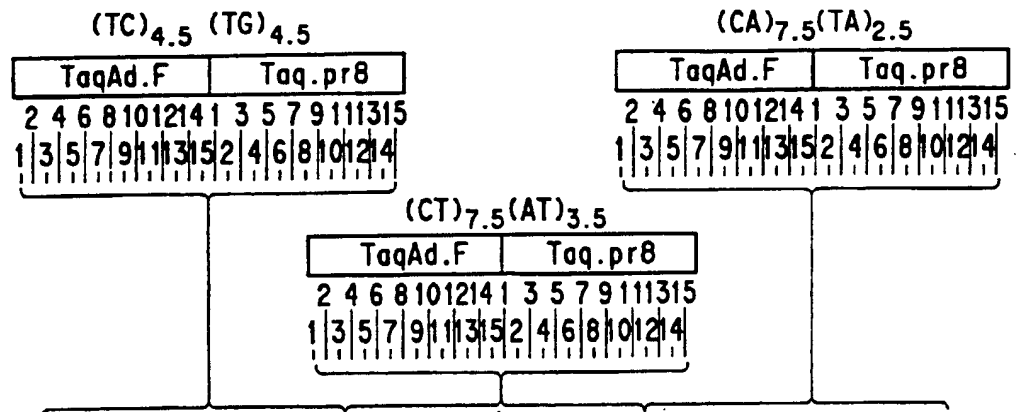


FIG. 3a

(TG) _{4.5} (TG) _{4.5}										(TC) _{4.5} (TC) _{4.5}																			
TaqAd.F					Taq.pr8					TaqAd.F					Taq.pr8														
2	4	6	8	10	12	14	1	3	5	7	9	11	13	15	2	4	6	8	10	12	14	1	3	5	7	9	11	13	15
1	3	5	7	9	11	13	15	2	4	6	8	10	12	14	1	3	5	7	9	11	13	15	2	4	6	8	10	12	14

FIG.3b



FIG.4

(CA)_{7.5} (TA)_{2.5} + Taq.pr6
F2 INDIVIDUALS

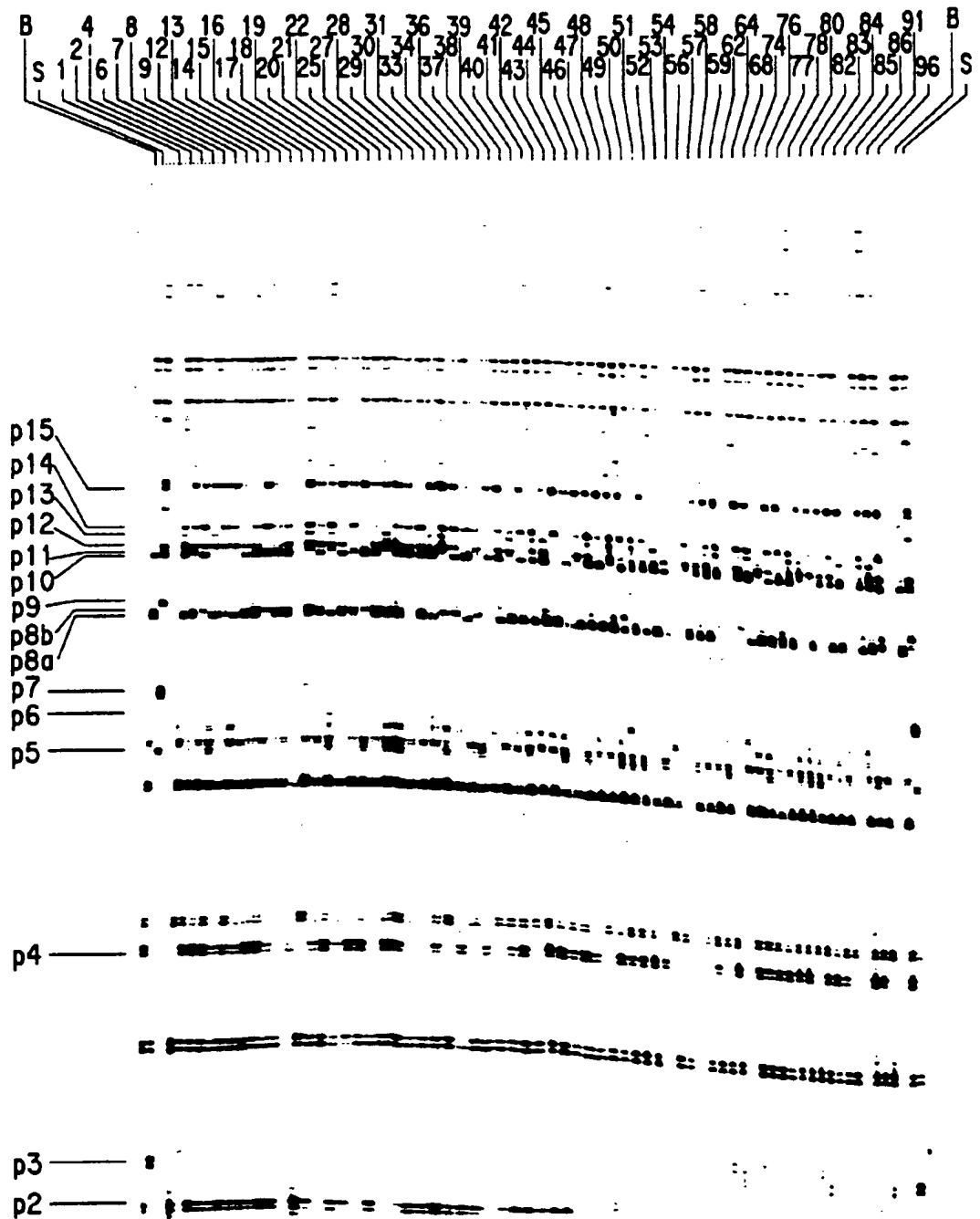


FIG.5a

FIG. 5b

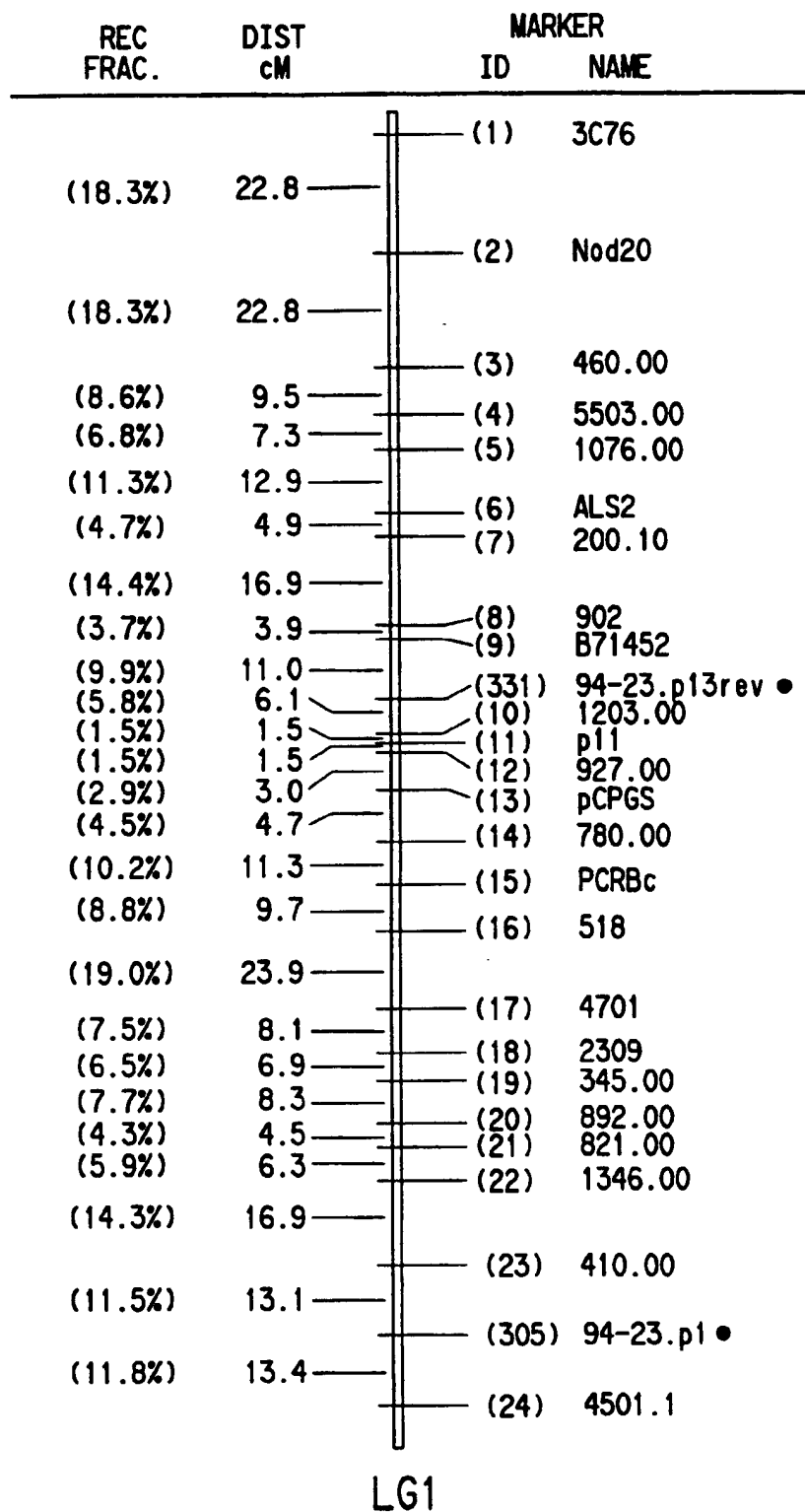


FIG. 5c

REC FRAC.	DIST cm	MARKER ID	NAME
(6.9%)	7.4	(159)	6011
(7.3%)	7.9	(160)	3807
(8.0%)	8.8	(161)	1029.00
		(162)	4510
(19.5%)	24.7		
		(163)	1201.00
(13.3%)	15.5		
		(164)	618.00
(20.2%)	25.9		
(7.8%)	8.5	(165)	6903.10
(5.7%)	6.0	(166)	1510.00
(7.5%)	8.1	(167)	4501.2
(7.7%)	8.4	(323)	94-23.p7
(2.4%)	2.4	(168)	OLEO
(8.3%)	9.1	(169)	1148.10
(8.3%)	9.0	(170)	3415
(5.4%)	5.7	(171)	6411.20
(3.0%)	3.0	(172)	7320.00
(4.9%)	5.1	(173)	1351.00
		(174)	2311

LG10

FIG. 5d

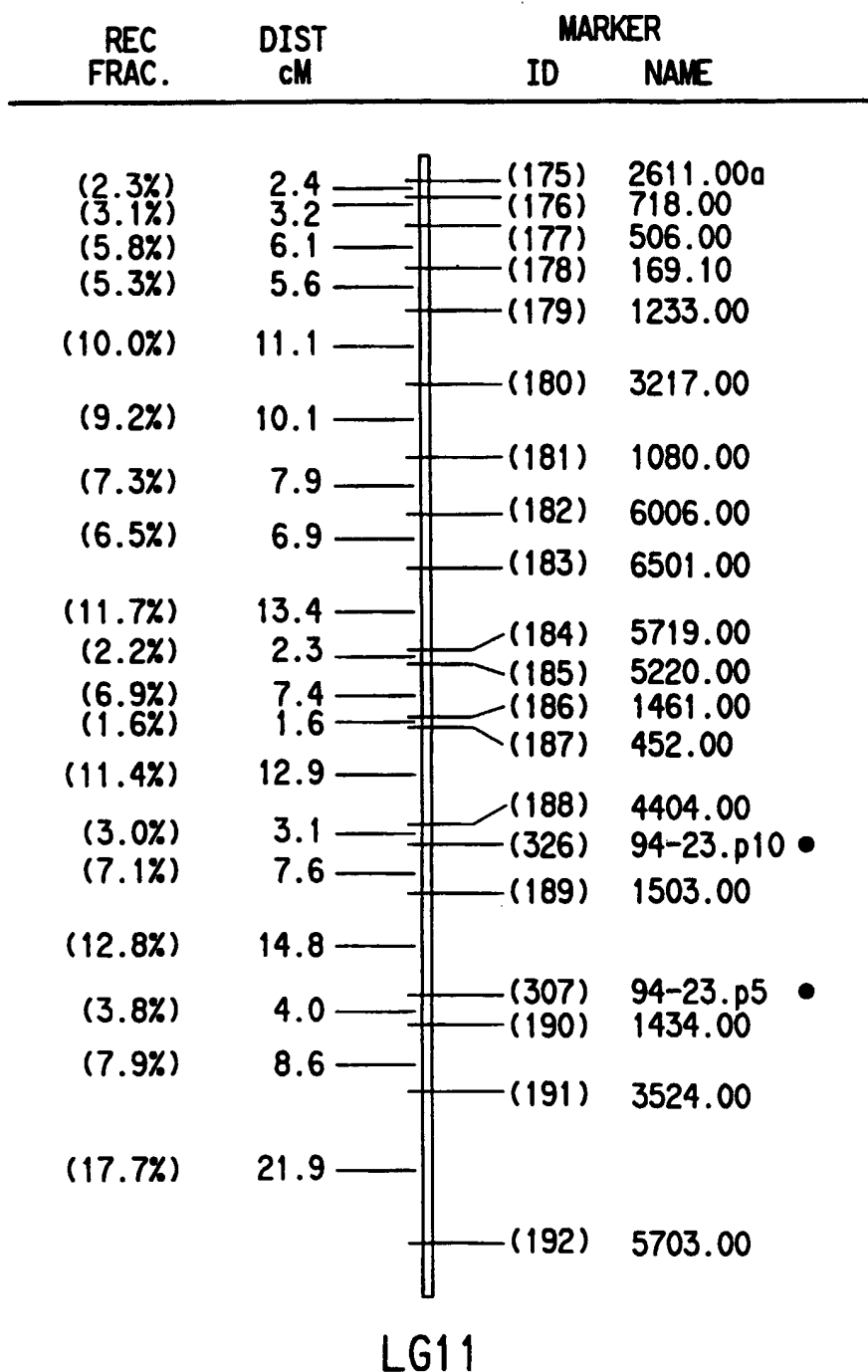


FIG. 5e

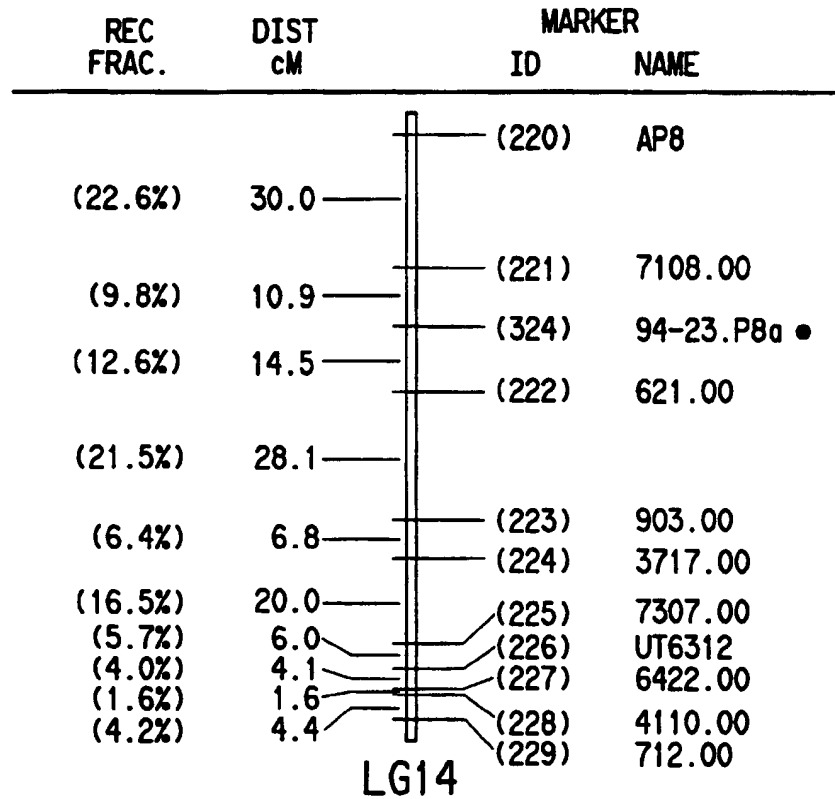
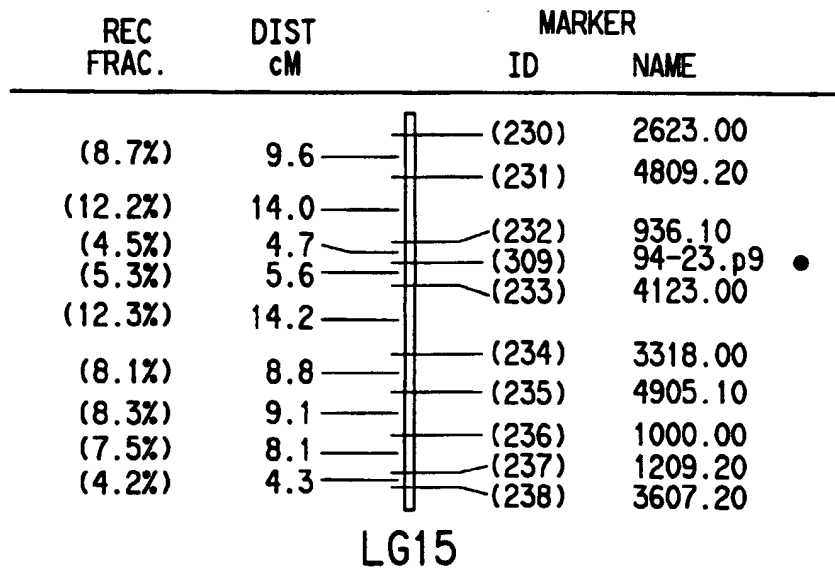


FIG. 5f



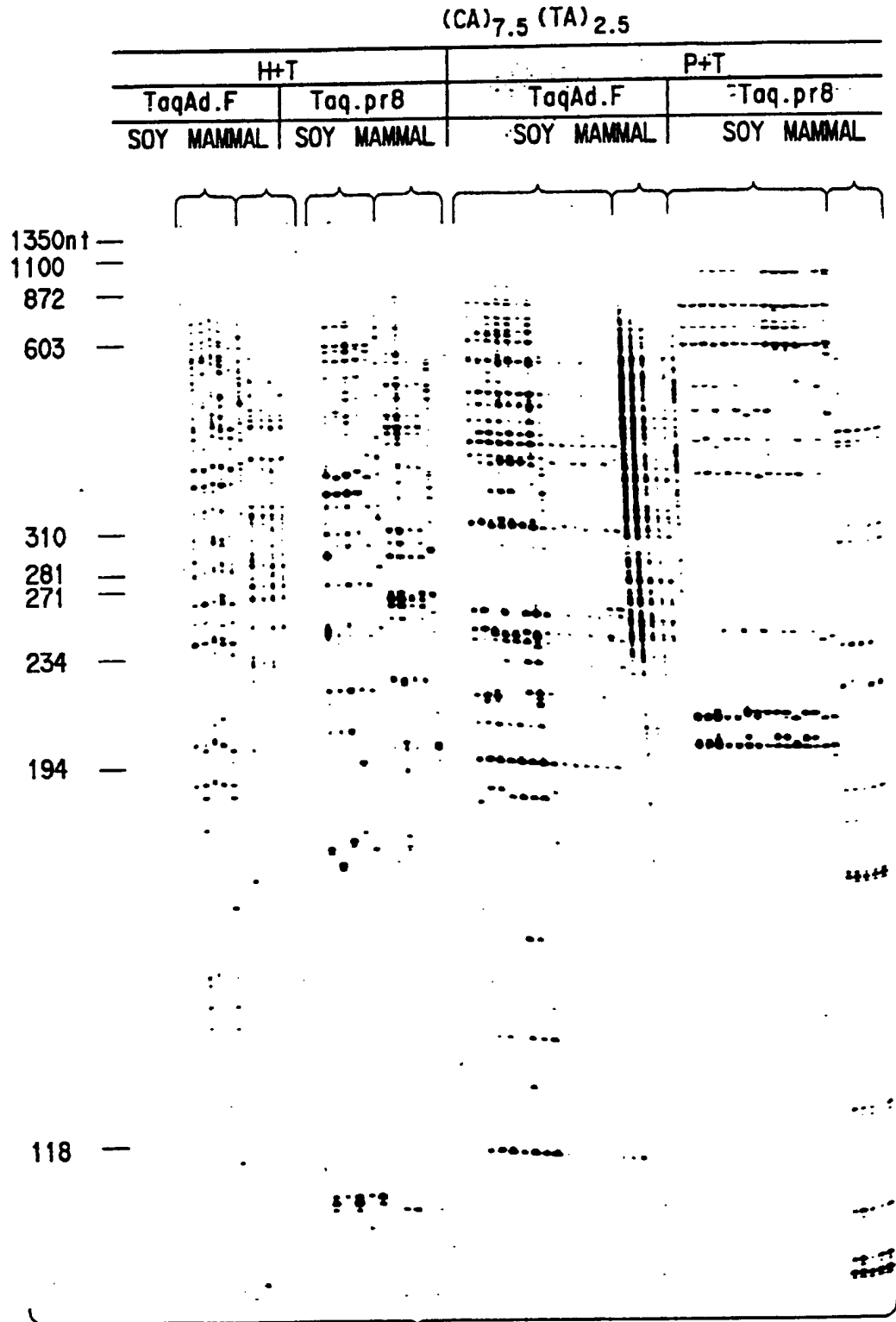


FIG. 6a

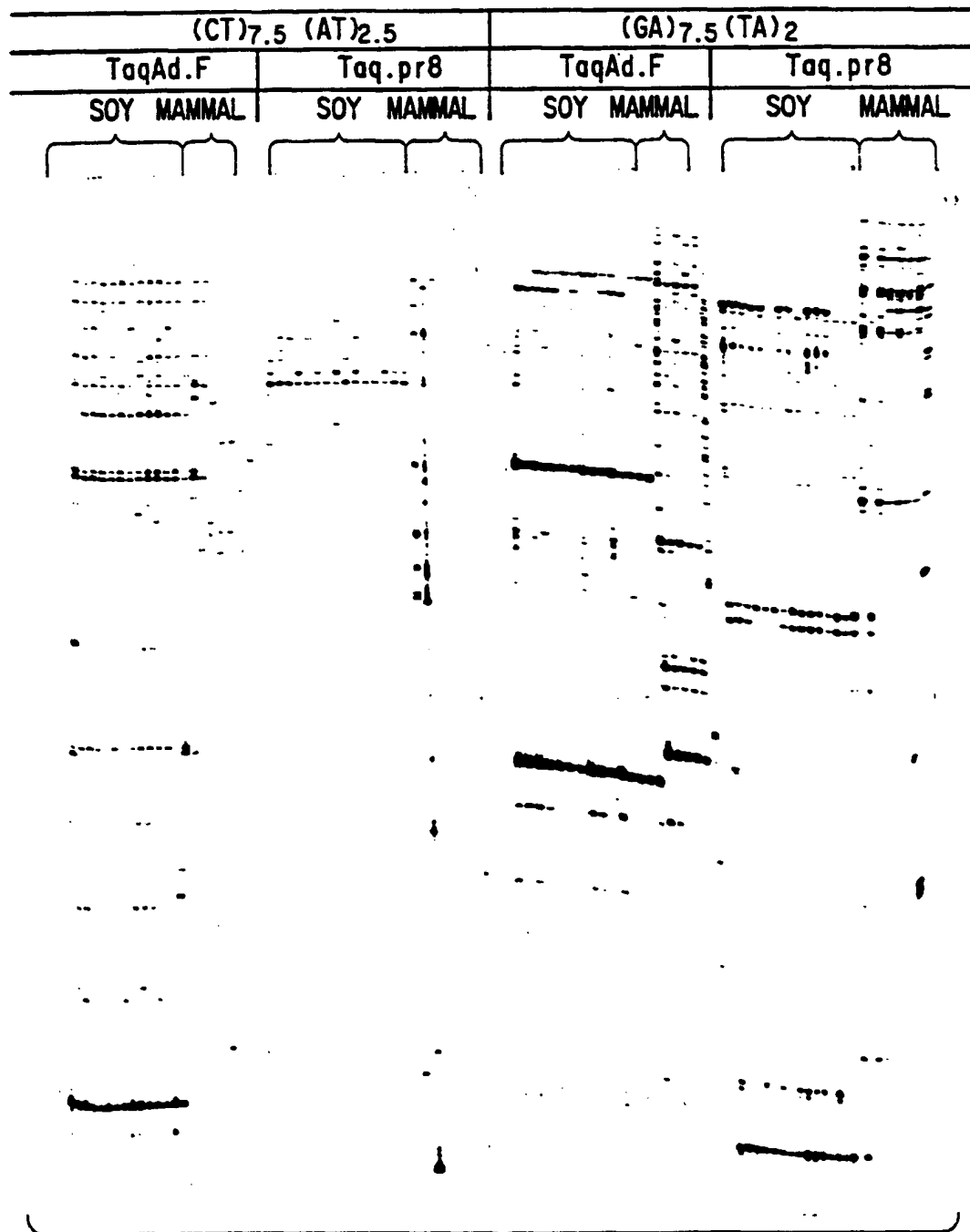


FIG. 6b

(CA)_{7.5} (TA)_{2.5} + Taq.pr6

2	4	6	9
1	3	5	7 8

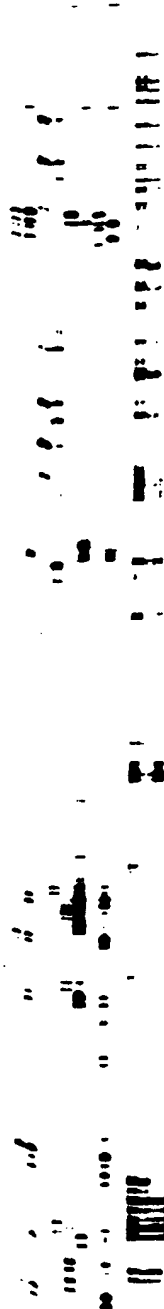


FIG. 6c

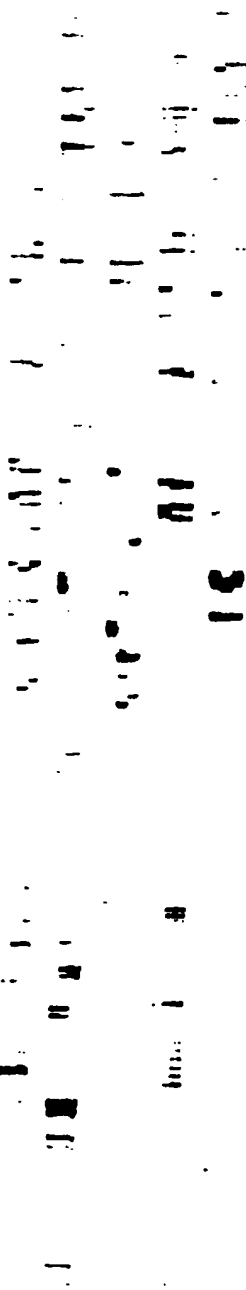
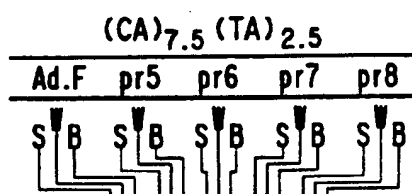


FIG.7

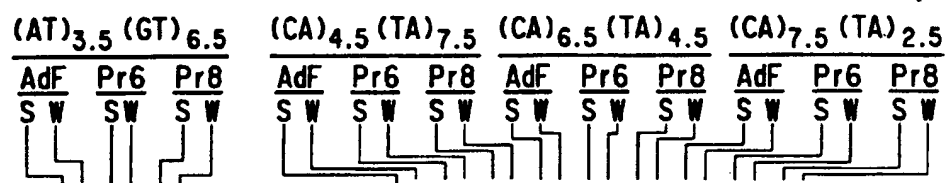


FIG.8

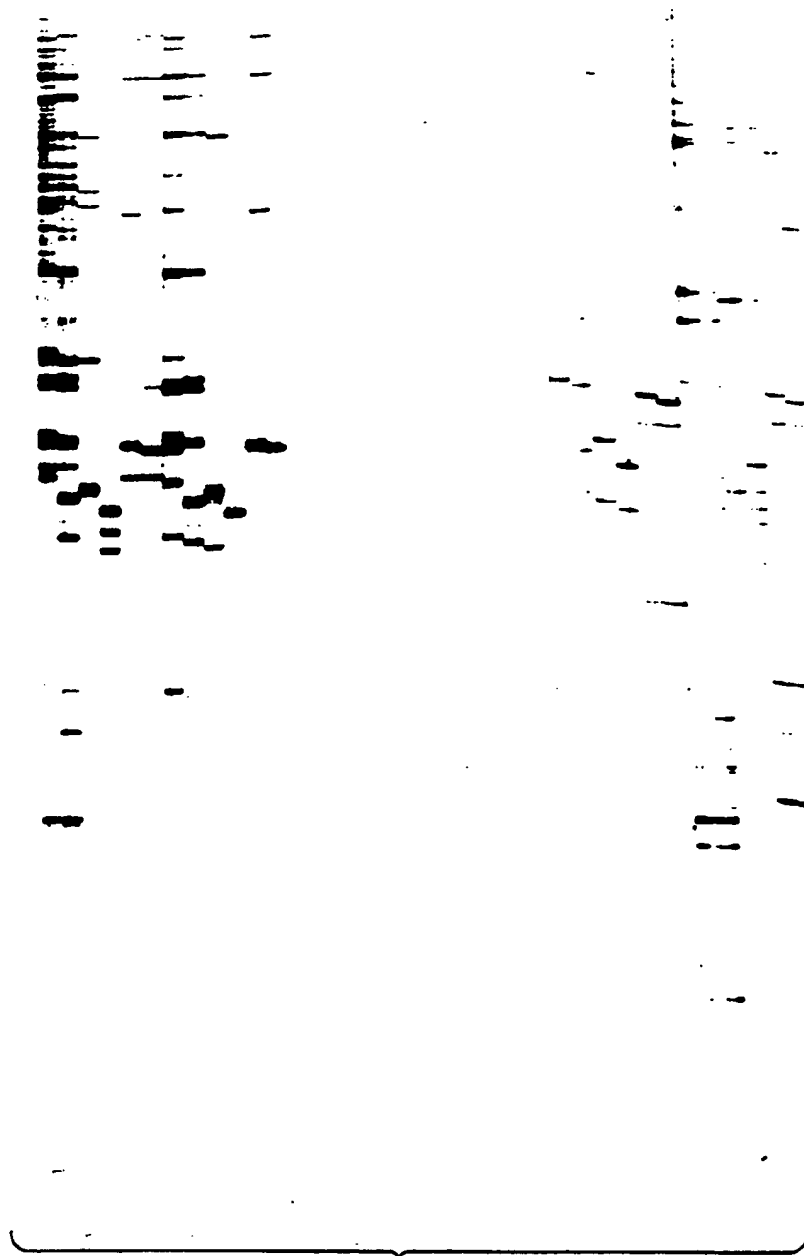
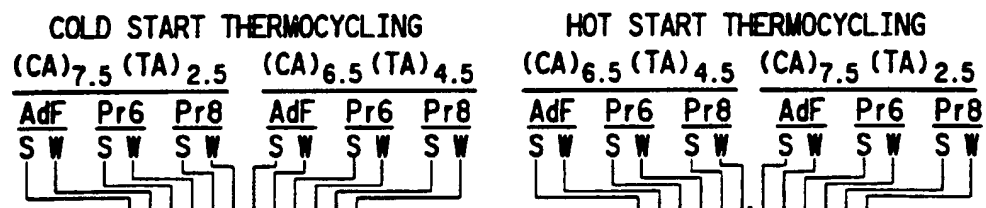


FIG. 9

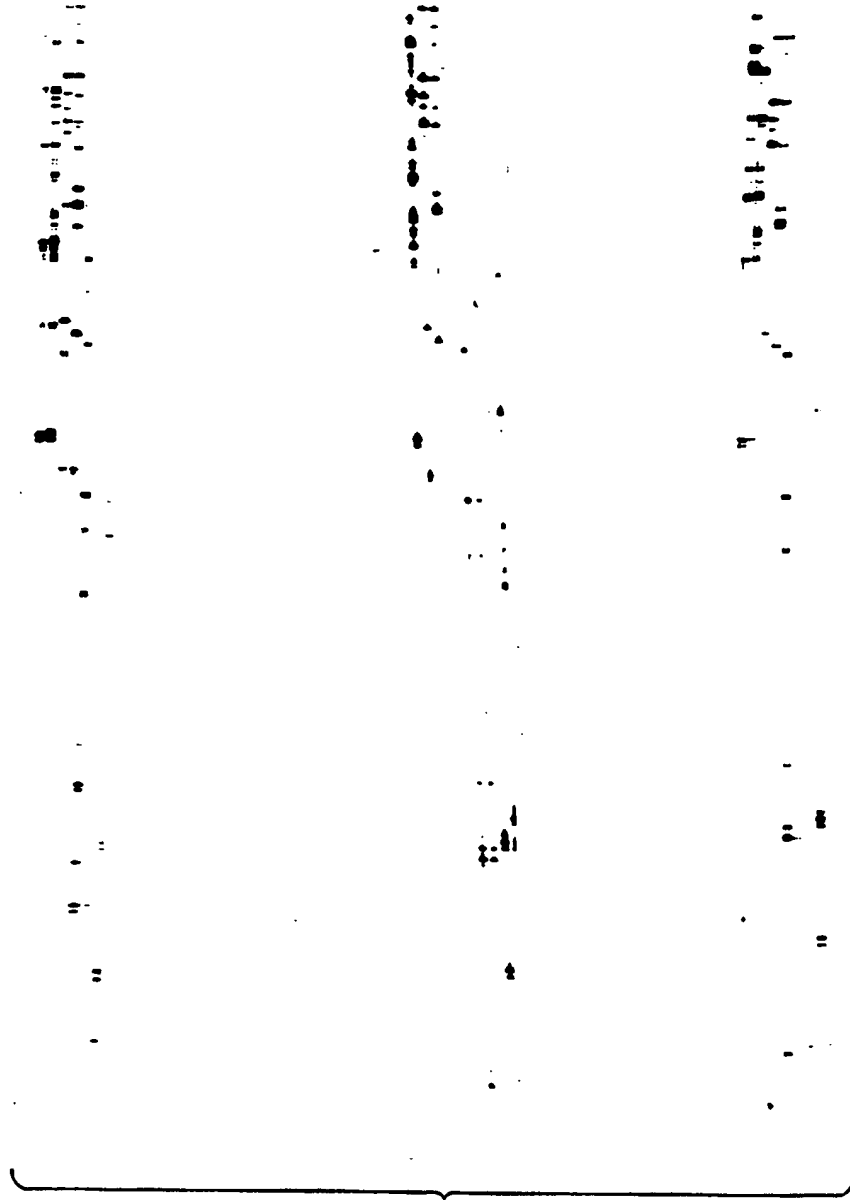
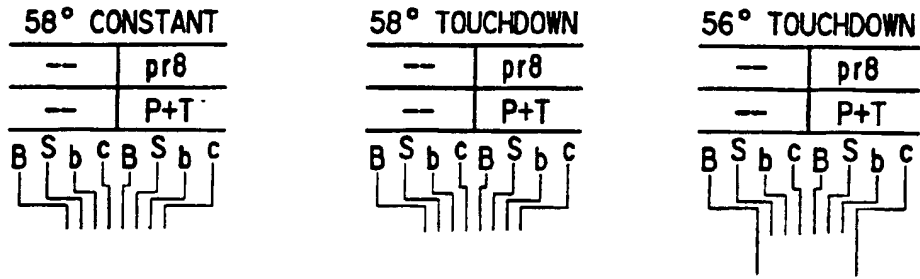


FIG. 10a

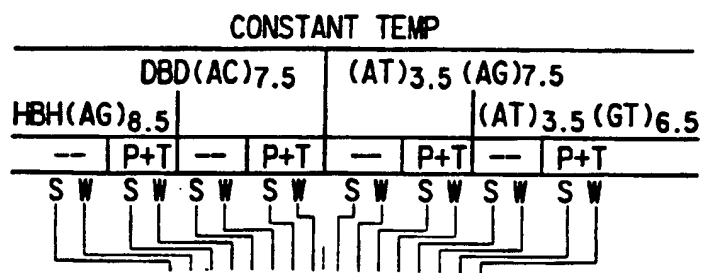


FIG. 10b

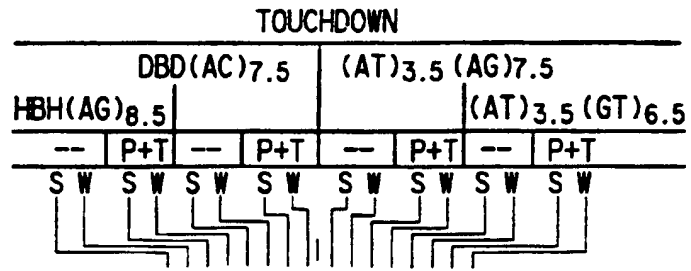
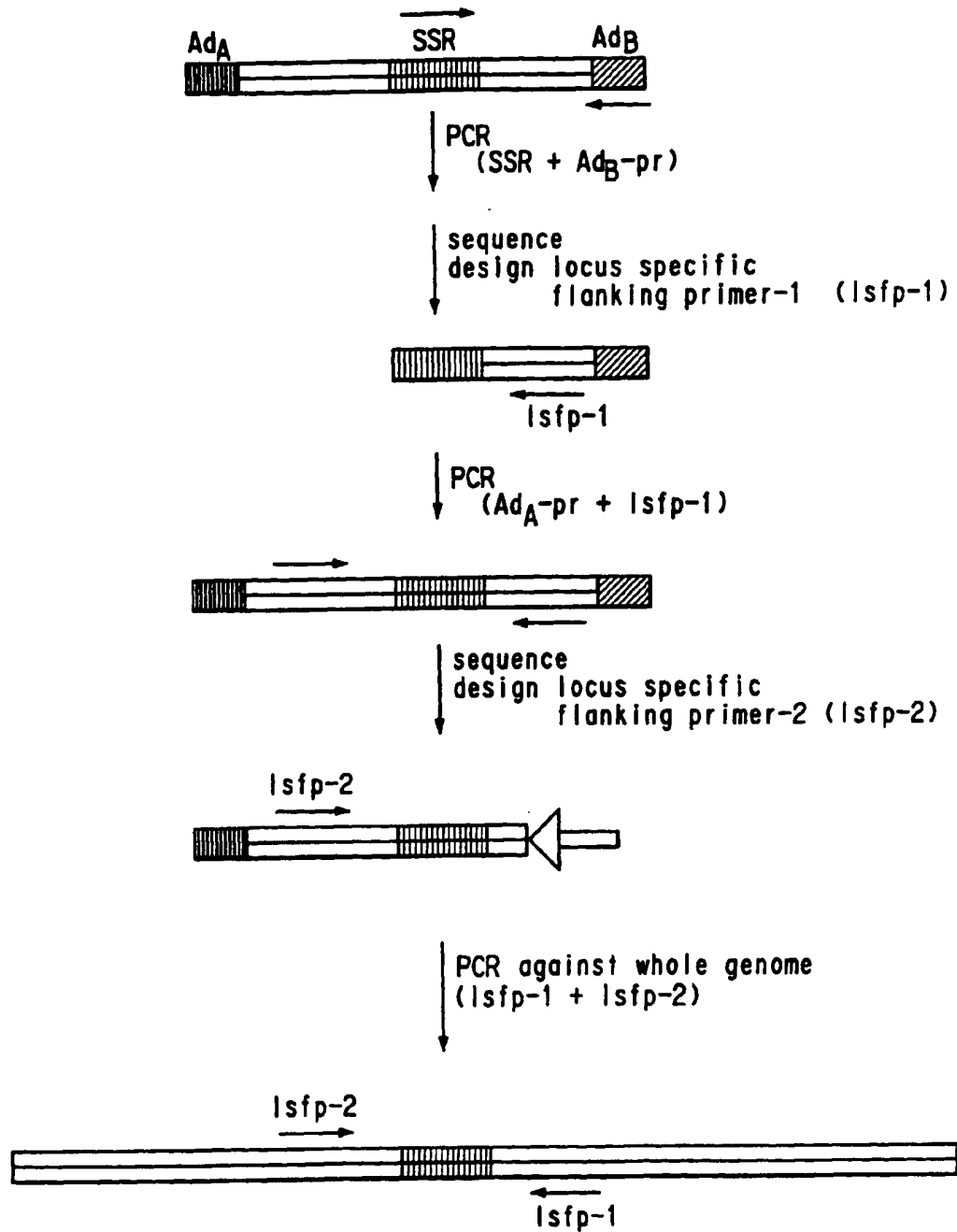


FIG. 10c

FIG. 11



PCTWORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : C12N 15/10, 9/02, 15/52	A1	(11) International Publication Number: WO 00/09682 (43) International Publication Date: 24 February 2000 (24.02.00)
(21) International Application Number: PCT/US99/18424 (22) International Filing Date: 12 August 1999 (12.08.99) (30) Priority Data: 60/096,271 12 August 1998 (12.08.98) US 60/130,810 23 April 1999 (23.04.99) US (71) Applicant (for all designated States except US): MAXYGEN, INC. [US/US]; 515 Galveston Drive, Redwood City, CA 94063 (US). (72) Inventors; and (75) Inventors/Applicants (for US only): AFFHOLTER, Joseph, A. [US/US]; 20520 Deepark Court, Saratoga, CA 95070 (US). DAVIS, Christopher [GB/US]; Apartment 103, 118 Church Street, San Francisco, CA 94114 (US). SELIFONOV, Sergey, A. [RU/US]; 2240 Homestead Court, Los Altos, CA 94024 (US). (74) Agents: MANN, Jeffry, S. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).		(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). Published <i>With international search report.</i> <i>Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>
(54) Title: DNA SHUFFLING OF MONOOXYGENASE GENES FOR PRODUCTION OF INDUSTRIAL CHEMICALS		
(57) Abstract <p>This invention provides improved monooxygenases, dehydrogenases, and transferases that are useful for the biocatalytic synthesis of compounds such as α-hydroxycarboxylic acids, and aryl- and alkyl-, hydroxy compounds. The polypeptides provided herein are improved in properties such as regioselectivity, enzymatic activity, stereospecificity, and the like. Methods for obtaining recombinant polynucleotides that encode these improved polypeptides are also provided, as are organisms that express the polypeptides and are thus useful for carrying out said biocatalytic syntheses. Also provided by the invention are methods for increasing said solvent resistance of organisms that are used in the synthetic methods.</p>		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

WO 00/09682

PCT/US99/18424

**DNA SHUFFLING OF MONOOXYGENASE GENES FOR
PRODUCTION OF INDUSTRIAL CHEMICALS**

5

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority under 35 U.S.C. §119(e) to U.S. Provisional Application Serial No. 60/096,271, and U.S. Provisional Application Serial No. 60/130,810, by Joseph A. Affholter, filed on August 12, 1998 and April 23, 1999, respectively. This application is related to the copending application titled DNA SHUFFLING OF
10 DIOXYGENASE GENES FOR PRODUCTION OF INDUSTRIAL CHEMICALS by Sergey A. Selifonov, Attorney Docket No. 018097-031100US, filed on an even day herewith. This application is also related to U.S. Provisional Application Serial No. 60/096,28, filed August 12, 1998, U.S. Provisional Application Serial No. 60/111,146, filed December 7, 1998, U.S. Provisional Application Serial No. 60/112,746, filed December 17,
15 1998. The disclosures of each the above-referenced applications are incorporated herein by reference in their entirety for all purposes.

FIELD OF THE INVENTION

This invention pertains to the shuffling of nucleic acids to achieve or enhance
20 industrial production of chemicals by monooxygenase genes.

BACKGROUND OF THE INVENTION

Organic acids, alcohols, aldehydes and epoxides are important classes of industrial chemicals. Typically, these products are generated by successive oxidation of
25 inexpensive, high volume saturated and unsaturated hydrocarbons (ethane, propane, butane, *etc.* and ethene, propene, butene, *etc.*) and simple aromatics such as benzene, ethyl benzene, naphthalene, styrene and toluene.

Monooxygenases (MOs) such as the P450 oxygenases, heme-dependent peroxidases, iron-sulfur MOs and quinone-dependent MOs typically catalyze limited
30 oxidation of these basic chemical building blocks. While potentially interesting from an industrial standpoint, these enzymes typically exhibit neither the physical robustness nor sufficient turnover numbers to make them usable as industrial catalysts. In addition, regeneration of a reduced heme is required following each catalytic turnover. Biologically,

the necessary heme reduction is mediated in the P450 family of enzymes by NAD(P)H, an expensive and impractical redox partner for most industrial chemistries.

Surprisingly, the present invention provides a method for providing enzymes with higher activity, high physical stability and robustness. Also surprisingly, the present invention provides a means of generating NADPH-independent monooxygenase activity in the presence of peroxide co-substrates (as well as other inexpensive cofactors) thereby solving each of the problems outlined above, as well as providing a variety of other features which will be apparent upon review.

SUMMARY OF THE INVENTION

In the present invention, DNA shuffling is used to generate new or improved monooxygenase genes. These monooxygenase genes are used to provide monooxygenase enzymes, especially for industrial processes. These new or improved genes have surprisingly superior properties as compared to naturally occurring monooxygenase genes.

In the methods for obtaining monooxygenase genes, a plurality of parental forms (homologs) of a selected nucleic acid are recombined. The selected nucleic acid is derived either from one or more parental nucleic acid(s) which encodes a monooxygenase enzyme, or a fragment thereof, or from a parental nucleic acid which does not encode monooxygenase, but which is a candidate for DNA shuffling to develop monooxygenase activity. The plurality of forms of the selected nucleic acid differ from each other in at least one (and typically two or more) nucleotides, and, upon recombination, provide a library of recombinant monooxygenase nucleic acids. The library can be an *in vitro* set of molecules, or present in cells, phage or the like. The library is screened to identify at least one recombinant monooxygenase nucleic acid that exhibits distinct or improved monooxygenase activity compared to the parental nucleic acid or nucleic acids.

Many formats for libraries of nucleic acids are known in the art and each of these formats is generally applicable to the libraries of the present invention. For example, basic texts generally disclosing library formats of use in this invention include Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual* (2nd ed. 1989); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); and *Current Protocols in Molecular Biology* (Ausubel *et al.*, eds., 1994)).

In a preferred embodiment, the starting DNA segments are first recombined by any of the formats described herein to generate a diverse library of recombinant DNA

segments. Such a library can vary widely in size from having fewer than 10 to more than 10^5 , 10^7 , or 10^9 members. In general, the starting segments and the recombinant libraries generated include full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. However, if this is not the case, the recombinant DNA segments in the library can be inserted into a common vector providing the missing sequences before performing screening/selection.

If the sequence recombination format employed is an *in vivo* format, the library of recombinant DNA segments generated already exists in a cell, which is usually the cell type in which expression of the enzyme with altered substrate specificity is desired. If sequence recombination is performed *in vitro*, the recombinant library is preferably introduced into the desired cell type before screening/selection. The members of the recombinant library can be linked to an episome or virus before introduction or can be introduced directly. In some embodiments of the invention, the library is amplified in a first host, and is then recovered from that host and introduced to a second host more amenable to expression, selection, or screening, or any other desirable parameter.

The manner in which the library is introduced into the cell type depends on the DNA-uptake characteristics of the cell type (e.g., having viral receptors, being capable of conjugation, or being naturally competent). If the cell type is not susceptible to natural and chemical-induced competence, but is susceptible to electroporation, one preferably employs electroporation. If the cell type is not susceptible to electroporation as well, one can employ biolistics. The biolistic PDS-1000 Gene Gun (Biorad, Hercules, Calif.) uses helium pressure to accelerate DNA-coated gold or tungsten microcarriers toward target cells. The process is applicable to a wide range of tissues, including plants, bacteria, fungi, algae, intact animal tissues, tissue culture cells, and animal embryos. One can employ electronic pulse delivery, which is essentially a mild electroporation format for live tissues in animals and patients. Zhao, *Advanced Drug Delivery Reviews* 17:257-262 (1995). Novel methods for making cells competent are described in co-pending application U.S. patent application Ser. No. 08/621,430, filed Mar. 25, 1996. After introduction of the library of recombinant DNA genes, the cells are optionally propagated to allow expression of genes to occur.

In selecting for monooxygenase activity, a candidate shuffled DNA can be tested for encoded monooxygenase activity in essentially any synthetic process. Common processes that can be screened include screening for alkane oxidation (e.g., hydroxylation, formation of ketones, aldehydes, etc.), screening for alkene epoxidation, aromatic

hydroxylation, N-dealkylation (*e.g.*, of alkylamines), S-dealkylation (*e.g.*, of reduced thio-organics), O-dealkylation (*e.g.*, of alkyl ethers), oxidation of aryloxy phenols, conversion of aldehydes to acids, alcohols to aldehydes or ketones, dehydrogenation, decarbonylation, oxidative dehalogenation of haloaromatics and halohydrocarbons, Baeyer-Villiger
5 monooxygenation, modification of cyclosporins, hydroxylation of mevastatin, hydroxylation of erythromycin, N-hydroxylation, sulfoxide formation, hydroxylation of fatty acids, hydroxylation of terpenes or oxygenation of sulfonylureas. Other oxidative transformations will be apparent to those of skill in the art.

Similarly, instead of, or in addition to, testing for an increase in
10 monooxygenase specific activity, it is also desirable to screen for shuffled nucleic acids which produce higher levels of monooxygenase nucleic acid or enhanced or reduced recombinant monooxygenase polypeptide expression or stability encoded by the recombinant monooxygenase nucleic acid.

A variety of screening methods can be used to screen a library, depending on
15 the monooxygenase activity for which the library is selected. By way of example, the library to be screened can be present in a population of cells. The library is selected by growing the cells in or on a medium comprising the chemical or compound to be oxidized or reduced and selecting for a detected physical difference between the oxidized or reduced form of the chemical or compound and the non-oxidized or reduced form of the chemical or compound,
20 either in the cell, or the extracellular medium.

Iterative selection for monooxygenase nucleic acids is also a feature of the invention. In these methods, a selected nucleic acid identified as encoding monooxygenase activity can be shuffled, either with the parental nucleic acids, or with other nucleic acids (*e.g.*, mutated forms of the selected nucleic acid) to produce a second shuffled library. The
25 second shuffled library is then selected for one or more form of monooxygenase activity, which can be the same or different than the monooxygenase activity previously selected. This process can be iteratively repeated as many times as desired, until a nucleic acid with optimized properties is obtained. If desired, any monooxygenase nucleic acid identified by any of the methods herein can be cloned and, optionally, expressed.

30 The invention also provides methods of increasing monooxygenase activity by whole genome shuffling. In these methods, a plurality of genomic nucleic acids are shuffled in a cell (in whole cell shuffling, entire genomes are shuffled, rather than specific sequences). The resulting shuffled nucleic acids are selected for one or more

monooxygenase traits. The genomic nucleic acids can be from a species or strain different from the cell in which monooxygenase activity is desired. Similarly, the shuffling reaction can be performed in cells using genomic DNA from the same or different species, or strains. Strains or enzymes exhibiting enhanced MO activity can be identified.

5 The distinct or improved monooxygenase activity encoded by a nucleic acid identified after shuffling can encode one or more of a variety of properties, including: an increased ability to chemically modify the monooxygenase target, an increase in the range of monooxygenase substrates which the distinct or improved nucleic acid operates on, an increase in the chemoselectivity of a polypeptide encoded by the nucleic acid, an increase in
10 the regioselectivity of a polypeptide encoded by the nucleic acid, an increase in the stereoselectivity of a polypeptide encoded by the nucleic acid, an increased expression level of a polypeptide encoded by the nucleic acid, a decrease in susceptibility of a polypeptide encoded by the nucleic acid to protease cleavage, a decrease in susceptibility of a
15 polypeptide encoded by the nucleic acid to high or low pH levels, a decrease in susceptibility of the protein encoded by the nucleic acid to high or low temperatures, a decrease in peroxide-mediated enzyme inactivation, a decrease in toxicity to a host cell of a polypeptide encoded by the selected nucleic acid, the ability to use low-cost reducing partners (rather than NAD(P)H), and a reduction in the sensitivity of the polypeptide and/or an organism expressing the polypeptide to inactivation by organic solvents and the feedstocks for and
20 products of the enzymatic oxidations, and

 The selected nucleic acids to be shuffled can be from any of a variety of sources, including synthetic or cloned DNAs. Exemplary targets for recombination include nucleic acids encoding P450 monooxygenases, nucleic acids encoding heme-dependent peroxidases, nucleic acids encoding iron sulfur monooxygenases, nucleic acids encoding
25 quinone-dependent monooxygenases, and the like. Typically, shuffled nucleic acids are cloned into expression vectors to achieve desired expression levels.

 In addition to shuffling monooxygenase nucleic acids, it is occasionally desirable to produce shuffled nucleic acids which produce oxidizing/reducing equivalents in forms other than O₂, H₂O₂ and NADPH, such as peroxides. Shuffled monooxygenase and
30 oxidase (H₂O₂) nucleic acids can be co-expressed in a single system to provide both monooxygenase activity and peroxide in a single system.

 One feature of the invention is production of libraries and shuffling mixtures for use in the methods as set forth above. For example, a phage display library comprising

shuffled forms of a nucleic acid is provided. Similarly, a shuffling mixture comprising at least three homologous DNAs, each of which is derived from a nucleic acid encoding a polypeptide or polypeptide fragment is provided. These polypeptides can be, for example, P450 monooxygenases, heme-dependent peroxidases, iron sulfur monooxygenases, quinone-dependent monooxygenases, and the like.

Isolated nucleic acids identified by selection of the libraries in the methods above are also a feature of the invention.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1. Schematic showing functional group insertion and modification using a monooxygenase.

Figure 2. Structures of exemplary feedstock olefinic compounds and structures of α -hydroxycarboxylic acids.

Figure 3. Enzymatic reaction schemes for multistep biochemical transformations of olefins to AHAs.

Figure 4. Enzymatic reaction schemes for converting free AHAs to ester derivatives.

Figure 5. Table of preferred MO reactions.

The absolute configuration of the chiral centers is not indicated in these Figures. The chiral centers of the chiral compounds can be R, S, or a mixture of these configurations.

DETAILED DESCRIPTION OF THE INVENTION AND THE PREFERRED EMBODIMENTS

Abbreviations

"AHA" refers to an α -hydroxycarboxylic acid.
"HCA" refers to a hydroxylated aromatic carboxylic acid
"MO" refers to a monooxygenase.

Definitions

Unless clearly indicated to the contrary, the following definitions supplement definitions of terms known in the art.

A "recombinant" nucleic acid is a nucleic acid produced by recombination between two or more nucleic acids, or any nucleic acid made by an *in vitro* or artificial

5

10

15

25

The terms "identical" or percent "identity," in the context of two or more nucleic acid or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the sequence comparison algorithms described below (or other algorithms available to persons of skill) or by visual inspection.

The phrase "substantially identical," in the context of two nucleic acids or polypeptides (*e.g.*, DNAs encoding a dioxygenase, or the amino acid sequence of the dioxygenase) refers to two or more sequences or subsequences that have at least about 60%, preferably 80%, most preferably 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. Such "substantially identical" sequences are typically considered to be homologous. Preferably, the "substantial identity" exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the sequences are substantially identical over at least about 150 residues, or over the full length of the two sequences to be compared.

For sequence comparison and homology determination, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally*, Ausubel *et al.*, *infra*).

One example of an algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in

Altschul *et al.*, *J. Mol. Biol.* 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which
5 either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be
10 increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved
15 value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For
20 amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin
25 & Altschul, *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to
30 the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

Another indication that two nucleic acid sequences are substantially identical/homologous is that the two molecules hybridize to each other under stringent conditions.

The phrase "hybridizing specifically to," refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions, including when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. "Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

"Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen LABORATORY TECHNIQUES IN BIOCHEMISTRY AND MOLECULAR BIOLOGY--HYBRIDIZATION WITH NUCLEIC ACID PROBES part I chapter 2 (1993) "Overview of principles of hybridization and the strategy of nucleic acid probe assays," Elsevier, New York. Generally, highly stringent hybridization and wash conditions are selected to be about 5 °C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Typically, under "stringent conditions" a probe will hybridize to its target subsequence, but not to unrelated sequences.

The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide with 1 mg of heparin at 42 °C, with the hybridization being carried out overnight. An example of highly stringent wash conditions is 0.15M NaCl at 72 °C for about 15 minutes. An example of stringent wash conditions is a 0.2x SSC wash at 65 °C for 15 minutes (*see*, Sambrook, *infra.*, for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example medium stringency wash for a duplex of, *e.g.*, more than 100 nucleotides, is 1x SSC at 45°C for 15 minutes. An example low stringency wash for a duplex of, *e.g.*, more than 100 nucleotides, is 4-6x SSC at 40 °C for 15 minutes. For short probes (*e.g.*, about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0

M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 30 °C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. Nucleic acids which do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, *e.g.*, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

A further indication that two nucleic acid sequences or polypeptides are substantially identical/homologous is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with, or specifically binds to, the polypeptide encoded by the second nucleic acid. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

"Conservatively modified variations" of a particular polynucleotide sequence refers to those polynucleotides that encode identical or essentially identical amino acid sequences, or where the polynucleotide does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given polypeptide. For instance, the codons CGU, CGC, CGA, CGG, AGA, and AGG all encode the amino acid arginine. Thus, at every position where an arginine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of "conservatively modified variations." Every polynucleotide sequence described herein which encodes a polypeptide also describes every possible silent variation, except where otherwise noted. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon for methionine) can be modified to yield a functionally identical molecule by standard techniques. Accordingly, each "silent variation" of a nucleic acid which encodes a polypeptide is implicit in each described sequence.

Furthermore, one of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the substitution of an

amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. The following five groups each contain amino acids that are conservative substitutions for one another:

- 5 Aliphatic: Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I);
 Aromatic: Phenylalanine (F), Tyrosine (Y), Tryptophan (W); Sulfur-containing:
 Methionine (M), Cysteine (C); Basic: Arginine (R), Lysine (K), Histidine (H); Acidic:
 Aspartic acid (D), Glutamic acid (E), Asparagine (N), Glutamine (Q). *See also*, Creighton
 (1984) *Proteins*, W.H. Freeman and Company. In addition, individual substitutions,
 deletions or additions which alter, add or delete a single amino acid or a small percentage of
10 amino acids in an encoded sequence are also "conservatively modified variations."
 Sequences that differ by conservative variations are generally homologous.

A "subsequence" refers to a sequence of nucleic acids or amino acids that comprise a part of a longer sequence of nucleic acids or amino acids (*e.g.*, polypeptide) respectively.

- 15 The term "gene" is used broadly to refer to any segment of DNA associated with expression of a given RNA or protein. Thus, genes include regions encoding expressed RNAs (which typically include polypeptide coding sequences) and, often, the regulatory sequences required for their expression. Genes can be obtained from a variety of sources, including cloning from a source of interest or synthesizing from known or predicted
20 sequence information, and may include sequences designed to have desired parameters.

The term "isolated", when applied to a nucleic acid or protein, denotes that the nucleic acid or protein is essentially free of other cellular components with which it is associated in the natural state.

- 25 The term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides which have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.*
30 degenerate codon substitutions) and complementary sequences and as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzner *et al.*, *Nucleic Acid Res.*

19:5081 (1991); Ohtsuka *et al.*, *J. Biol. Chem.* 260:2605-2608 (1985); Cassol *et al.* (1992); Rossolini *et al.*, *Mol. Cell. Probes* 8:91-98 (1994)). The term nucleic acid is generic to the terms "gene", "DNA", "cDNA", "oligonucleotide", "RNA", "mRNA", "polynucleotide" and the like.

5 "Nucleic acid derived from a gene" refers to a nucleic acid for whose synthesis the gene, or a subsequence thereof, has ultimately served as a template. Thus, an mRNA, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, *etc.*, are all derived from the gene and detection of such derived products is indicative of the presence
10 and/or abundance of the original gene and/or gene transcript in a sample.

A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it increases the transcription of the coding sequence.

A "recombinant expression cassette" or simply an "expression cassette" is a
15 nucleic acid construct, generated recombinantly or synthetically, with nucleic acid elements that are capable of effecting expression of a structural gene in hosts compatible with such sequences. Expression cassettes include at least promoters and optionally, transcription termination signals. Typically, the recombinant expression cassette includes a nucleic acid to be transcribed (*e.g.*, a nucleic acid encoding a desired polypeptide), and a promoter.
20 Additional factors necessary or helpful in effecting expression may also be used as described herein. For example, an expression cassette can also include nucleotide sequences that encode a signal sequence that directs secretion of an expressed protein from the host cell. Transcription termination signals, enhancers, and other nucleic acid sequences that influence gene expression, can also be included in an expression cassette.

25 The term "NAD(P)H" is used herein to refer to the reducing agents, NADH and NADPH.

"Regioselectivity" is used herein to refer to the ability to discriminate between different positions of the monooxygenase target.

"Chemoselectivity" is used herein to refer to the ability to discriminate
30 between two or more potential sites of action in the monooxygenase target (*e.g.* alkyl hydroxylation in the presence of an epoxide and the like).

"Stereoselectivity" is used herein to refer to the ability to discriminate between enantiomeric sites in the monooxygenase target.

"Alkyl" refers to straight- and branched-chain, saturated and unsaturated hydrocarbons. "Lower alkyl", as used herein, refers to "alkyl" groups having from about 1 to about 6 carbon atoms.

"Substituted alkyl" refers to alkyl as just described including one or more functional groups such as lower alkyl, aryl, acyl, halogen (*i.e.*, alkylhalos, *e.g.*, CF₃), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, aryloxy, aryloxyalkyl, mercapto, both saturated and unsaturated cyclic hydrocarbons, heterocycles and the like. These groups may be attached to any carbon of the alkyl moiety.

The term "aryl" is used herein to refer to an aromatic substituent which may be a single aromatic ring or multiple aromatic rings which are fused together, linked covalently, or linked to a common group such as a methylene or ethylene moiety. The common linking group may also be a carbonyl as in benzophenone. The aromatic ring(s) may include phenyl, naphthyl, biphenyl, diphenylmethyl and benzophenone among others. The term "aryl" encompasses "arylalkyl."

The term "alkylarene" is used herein to refer to a subset of "aryl" in which the aryl group is substituted with an alkyl group as defined herein.

"Substituted aryl" refers to aryl as just described including one or more functional groups such as lower alkyl, acyl, halogen, alkylhalos (*e.g.* CF₃), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, mercapto and both saturated and unsaturated cyclic hydrocarbons which are fused to the aromatic ring(s), linked covalently or linked to a common group such as a methylene or ethylene moiety. The linking group may also be a carbonyl such as in cyclohexyl phenyl ketone. The term "substituted aryl" encompasses "substituted arylalkyl."

The term "acyl" is used to describe a ketone substituent, —C(O)R, wherein R is alkyl or substituted alkyl, aryl or substituted aryl as defined herein.

The term "halogen" is used herein to refer to fluorine, bromine, chlorine and iodine atoms.

The term "hydroxy" is used herein to refer to the group —OH.

The term "amino" is used to describe primary amines, R—NH₂, wherein R is alkyl or substituted alkyl, aryl or substituted aryl as defined herein.

The term "alkoxy" is used herein to refer to the —OR group, wherein R is a lower alkyl, substituted lower alkyl, aryl, substituted aryl, arylalkyl or substituted arylalkyl wherein the alkyl, aryl, substituted aryl, arylalkyl and substituted arylalkyl groups are as

described herein. Suitable alkoxy radicals include, for example, methoxy, ethoxy, phenoxy, substituted phenoxy, benzyloxy, phenethyloxy, t-butoxy, *etc.*

The term "alkylamino" denotes secondary and tertiary amines wherein the alkyl groups may be either the same or different and may consist of straight or branched, saturated or unsaturated hydrocarbons.

The term "unsaturated cyclic hydrocarbon" is used to describe a non-aromatic group with at least one double bond, such as cyclopentene, cyclohexene, *etc.* and substituted analogues thereof.

The term "heteroaryl" as used herein refers to aromatic rings in which one or more carbon atoms of the aromatic ring(s) are substituted by a heteroatom such as nitrogen, oxygen or sulfur. Heteroaryl refers to structures which may be a single aromatic ring, multiple aromatic ring(s), or one or more aromatic rings coupled to one or more non-aromatic ring(s). In structures having multiple rings, the rings can be fused together, linked covalently, or linked to a common group such as a methylene or ethylene moiety. The common linking group may also be a carbonyl as in phenyl pyridyl ketone. As used herein, rings such as thiophene, pyridine, isoxazole, phthalimide, pyrazole, indole, furan, *etc.* or benzo-fused analogues of these rings are defined by the term "heteroaryl."

"Alkylheteroaryl" defines a subset of "heteroaryl" substituted with an alkyl group, as defined herein.

"Substituted heteroaryl" refers to heteroaryl as just described wherein the heteroaryl nucleus is substituted with one or more functional groups such as lower alkyl, acyl, halogen, alkylhalos (*e.g.* CF₃), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, mercapto, *etc.* Thus, substituted analogues of heteroaromatic rings such as thiophene, pyridine, isoxazole, phthalimide, pyrazole, indole, furan, *etc.* or benzo-fused analogues of these rings are defined by the term "substituted heteroaryl."

The term "heterocyclic" is used herein to describe a saturated or unsaturated non-aromatic group having a single ring or multiple condensed rings from about 1 to about 12 carbon atoms and from about 1 to about 4 heteroatoms selected from nitrogen, sulfur or oxygen within the ring. Such heterocycles are, for example, tetrahydrofuran, morpholine, piperidine, pyrrolidine, *etc.*

The term "substituted heterocyclic" as used herein describes a subset of "heterocyclic" wherein the heterocycle nucleus is substituted with one or more functional

groups such as lower alkyl, acyl, halogen, alkylhalos (*e.g.* CF₃), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, mercapto, *etc.*

The term "alkylheterocyclyl" defines a subset of "heterocyclic" substituted with an alkyl group, as defined herein.

5 The term "substituted heterocyclicalkyl" defines a subset of "heterocyclic alkyl" wherein the heterocyclic nucleus is substituted with one or more functional groups such as lower alkyl, acyl, halogen, alkylhalos (*e.g.* CF₃), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, mercapto, *etc.*

10 Introductio..

 This invention describes the generation of evolved monooxygenases with enhanced performance for use in the production of chemicals of industrial interest using any of a variety of shuffling techniques, including, for example, gene, family and whole genome shuffling as described herein. In this invention, shuffling is used to enhance properties of
15 monooxygenases, such as forward rate kinetics, substrate specificity, regioselectivity, chemoselectivity, stereoselectivity and affinity and also to decrease susceptibility of monooxygenases to reversible inhibitors and inactivation by solvents, starting materials and reaction products and intermediates generated during the catalytic cycle.

 While much of the discussion below deals explicitly with P450
20 monooxygenases, this is for clarity of illustration. The discussion is representative of the chemistries and improvements which can be made to other useful monooxygenases, such as the structurally and functionally similar peroxidases and chlorperoxidases, as well as to the structurally unrelated iron-sulfur methane monooxygenases and other enzymes noted herein using the gene and family shuffling methodologies described.

25 In a first aspect, the present invention provides a method for obtaining a nucleic acid that encodes an improved polypeptide possessing monooxygenase activity. The improved polypeptide has at least one property improved over a naturally occurring monooxygenase polypeptide. The method includes: (a) creating a library of recombinant polynucleotides encoding a recombinant monooxygenase polypeptide; and (b) screening the
30 library to identify a recombinant polynucleotide that encodes an improved recombinant monooxygenase polypeptide that has at least one property improved over the naturally occurring polypeptide. Also provided are nucleic acids produced by this method that encode

a monooxygenase polypeptide having at least one property improved over a naturally occurring monooxygenase polypeptide.

In a preferred embodiment, the nucleic acid libraries of the invention are constructed by a method that includes shuffling a plurality of parental polynucleotides to produce one or more recombinant monooxygenase polynucleotide encoding the improved property. In another preferred embodiment, the polynucleotides are homologous. A detailed description of shuffling techniques is provided in Part A, hereinbelow.

In another embodiment, at least one of the parental polynucleotides is selected from polynucleotides that encode at least one monooxygenase activity and those that do not encode at least one monooxygenase activity. Typically, the parental monooxygenase polynucleotide encodes a complete polypeptide or a polypeptide fragment selected from an arene monooxygenase or fragments thereof.

In a preferred embodiment, the monooxygenase activity is a member selected from alkane oxidation (*e.g.*, hydroxylation, formation of ketones, aldehydes, *etc.*), alkene epoxidation, aromatic hydroxylation, N-dealkylation (*e.g.*, of alkylamines), S-dealkylation (*e.g.*, of reduced thio-organics), O-dealkylation (*e.g.*, of alkyl ethers), oxidation of aryloxy phenols, conversion of aldehydes to acids, alcohols to aldehydes or ketones, dehydrogenation, decarbonylation, oxidative dehalogenation of haloaromatics and halohydrocarbons, Baeyer-Villiger monooxygenation, modification of cyclosporins, hydroxylation of mevastatin, hydroxylation of erythromycin, hydroxylations of fatty acids, hydroxylation/epoxidation of terpenes, N-hydroxylation, sulfoxide formation, or oxygenation of sulfonylureas. Other oxidative transformations will be apparent to those of skill in the art.

The invention provides significant advantages over previously used methods for optimization of monooxygenase genes. For example, DNA shuffling can result in optimization of a desirable property even in the absence of a detailed understanding of the mechanism by which the particular property is mediated. In addition, entirely new properties can be obtained upon shuffling of DNAs, *i.e.*, shuffled DNAs can encode polypeptides or RNAs with properties entirely absent in the parental DNAs which are shuffled.

The properties or characteristics that can be acquired or improved vary widely, and depend on the choice of substrate. For example, for monooxygenase genes, properties that one can improve include, but are not limited to, increased range of

monooxygenases activity encoded by a particular gene, increased potency against a monooxygenase target, increased regioselectivity of action against a monooxygenase target, increased chemoselectivity of action against a monooxygenase target, increased stereoselectivity of action against a monooxygenase target, increased expression level of the monooxygenase gene, increased tolerance of the protein encoded by the monooxygenase gene to protease degradation (or other natural protein or RNA degradative processes), increased monooxygenase activity ranges for conditions such as heat, cold, low or high pH, reduced toxicity to the host cell, and increased resistance of the polypeptide and/or the organism expressing the polypeptide to organic solvents, and reaction feedstocks, intermediates and products.

The targets for modification vary in different applications, as does the property sought to be acquired or improved. Examples of candidate targets for acquisition of a property or improvement in a property include genes that encode proteins which have enzymatic or other activities useful in monooxygenase reactions.

The methods typically use at least two variant forms of a starting target. The variant forms of candidate substrates can show substantial sequence or secondary structural similarity with each other, but they should also differ in at least one and preferably at least two positions.

The initial diversity between forms can be the result of natural variation, *e.g.*, the different variant forms (homologs) are obtained from different individuals or strains of an organism, or constitute related sequences from the same organism (*e.g.*, allelic variations), or constitute homologs from different organisms (interspecific variants). Alternatively, initial diversity can be induced, *e.g.*, the variant forms can be generated by error-prone transcription, such as an error-prone PCR or use of a polymerase which lacks proof-reading activity (*see*, Liao, *Gene* 88:107-111 (1990)), of the first variant form, or, by replication of the first form in a mutator strain (mutator host cells are discussed in further detail below, and are generally well known). Alternatively, initial diversity can be generated by the creation of chimeric nucleic acids. The initial diversity between substrates is greatly augmented in subsequent steps of recombination for library generation.

A mutator strain can include any mutants in any organism impaired in the functions of mismatch repair. These include mutant gene products of *mutS*, *mutT*, *mutH*, *mutL*, *ovrD*, *dcm*, *vsr*, *umuC*, *umuD*, *sbcB*, *recJ*, *etc.* The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small

molecule or an expressed antisense RNA, or other techniques. Impairment can be of the genes noted, or of homologous genes in any organism.

Therefore, in carrying out the practice of the present invention, at least two variant forms of a nucleic acid which can confer monooxygenase activity are recombined to produce a library of recombinant monooxygenase genes. The library is then screened to identify at least one recombinant monooxygenase gene that is optimized for the particular property or properties of interest.

The parental polynucleotides can be shuffled in substantially any cell type, including prokaryotes, eukaryotes, yeast, bacteria and fungi. In a preferred embodiment, the one or more recombinant monooxygenase nucleic acid is present in one or more bacterial, yeast, or fungal cells and the method includes: pooling multiple separate monooxygenase nucleic acids; screening the resulting pooled monooxygenase nucleic acids to identify a distinct or improved recombinant monooxygenase nucleic acids that exhibit distinct or improved monooxygenase activity compared to a non-recombinant monooxygenase activity nucleic acid; and cloning the distinct or improved recombinant nucleic acid.

Often, improvements are achieved after one round of recombination and selection. However, recursive sequence recombination can be employed to achieve still further improvements in a desired property, or to bring about new (or "distinct") properties. Recursive sequence recombination entails successive cycles of recombination to generate molecular diversity. That is, one creates a family of nucleic acid molecules showing some sequence identity to each other but differing in the presence of mutations. In any given cycle, recombination can occur *in vivo* or *in vitro*, intracellularly or extracellularly. Furthermore, diversity resulting from recombination can be augmented in any cycle by applying prior methods of mutagenesis (*e.g.*, error-prone PCR or cassette mutagenesis) to either the substrates or products for recombination.

A recombination cycle is usually followed by at least one cycle of screening or selection for molecules having a desired property or characteristic. If a recombination cycle is performed *in vitro*, the products of recombination, *i.e.*, recombinant segments, are sometimes introduced into cells before the screening step. Recombinant segments can also be linked to an appropriate vector or other regulatory sequences before screening. Alternatively, products of recombination generated *in vitro* are sometimes packaged in viruses (*e.g.*, bacteriophage) before screening. If recombination is performed *in vivo*, recombination products can sometimes be screened in the cells in which recombination

occurred. In other applications, recombinant segments are extracted from the cells, and optionally packaged as viruses, before screening.

The nature of screening or selection depends on what property or characteristic is to be acquired or the property or characteristic for which improvement is sought, and many examples are discussed below. It is not usually necessary to understand the molecular basis by which particular products of recombination (recombinant segments) have acquired new or improved properties or characteristics relative to the starting substrates. For example, a monooxygenase gene can have many component sequences each having a different intended role (e.g., coding sequence, regulatory sequences, targeting sequences, stability-conferring sequences, subunit sequences and sequences affecting integration). Each of these component sequences can be varied and recombined simultaneously. Screening/selection can then be performed, for example, for recombinant segments that have increased ability to confer monooxygenase activity upon a cell without the need to attribute such improvement to any of the individual component sequences of the vector.

Depending on the particular screening protocol used for a desired property, initial round(s) of screening can sometimes be performed using bacterial cells due to high transfection efficiencies and ease of culture. However, for eukaryotic monooxygenases such as eukaryotic arene monooxygenases, bacterial expression is often not practical, and yeast, fungal or other eukaryotic systems are used for library expression and screening. Similarly other types of screening which are not amenable to screening in bacterial or simple eukaryotic library cells, are performed in cells selected for use in an environment close to that of their intended use. Final rounds of screening can be performed in the precise cell type of intended use.

If further improvement in a property is desired, at least one and usually a collection of recombinant segments surviving a first round of screening/selection are subject to a further round of recombination. These recombinant segments can be recombined with each other or with exogenous segments representing the original substrates or further variants thereof. Again, recombination can proceed *in vitro* or *in vivo*. If the previous screening step identifies desired recombinant segments as components of cells, the components can be subjected to further recombination *in vivo*, or can be subjected to further recombination *in vitro*, or can be isolated before performing a round of *in vitro* recombination. Conversely, if the previous screening step identifies desired recombinant

segments in naked form or as components of viruses, these segments can be introduced into cells to perform a round of *in vivo* recombination. The second round of recombination, irrespective how performed, generates further recombinant segments which encompass additional diversity than is present in recombinant segments resulting from previous rounds.

- 5 The second round of recombination can be followed by a further round of screening/selection according to the principles discussed above for the first round. The stringency of screening/selection can be increased between rounds. Also, the nature of the screen and the property being screened for can vary between rounds if improvement in more than one property is desired or if acquiring more than one new property is desired.
- 10 Additional rounds of recombination and screening can then be performed until the recombinant segments have sufficiently evolved to acquire the desired new or improved property or function.

- In a preferred embodiment, the invention provides a recursive method for making a nucleic acid encoding a specific monooxygenase activity. In this method, the
- 15 parental nucleic acids are shuffled in a plurality of cells and the method optionally further includes one or more of: (a) recombining DNA from the plurality of cells that display monooxygenase activity with a library of DNA fragments, at least one of which undergoes recombination with a segment in a cellular DNA present in the cells to produce recombined cells, or recombining DNA between the plurality of cells that display monooxygenase
- 20 activity to produce cells with modified monooxygenase activity; (b) recombining and screening the recombined or modified cells to produce further recombined cells that have evolved additionally modified monooxygenase activity; and, (c) repeating (a) or (b) until the further recombined cells have acquired a desired monooxygenase activity.

- In another preferred embodiment, the invention provides a method for making
- 25 a nucleic acid encoding a specific monooxygenase activity. This method includes: (a) recombining at least one distinct or improved recombinant nucleic acid with a further monooxygenase activity nucleic acid, which further nucleic acid is the same or different from one or more of the plurality of parental nucleic acids to produce a library of recombinant monooxygenase nucleic acids; (b) screening the library to identify at least one
- 30 further distinct or improved recombinant monooxygenase nucleic acid that exhibits a further improvement or distinct property compared to the plurality of parental nucleic acids; and, optionally; (c) repeating (a) and (b) until the resulting further distinct or improved

recombinant nucleic acid shows an additionally distinct or improved monooxygenase property.

The practice of this invention involves the construction of recombinant nucleic acids and the expression of genes in transfected host cells. Molecular cloning techniques to achieve these ends are known in the art. A wide variety of cloning and *in vitro* amplification methods suitable for the construction of recombinant nucleic acids such as expression vectors are well-known to persons of skill. General texts which describe molecular biological techniques useful herein, including mutagenesis, include Berger and Kimmel, GUIDE TO MOLECULAR CLONING TECHNIQUES, METHODS IN ENZYMOLOGY, volume 152, Academic Press, Inc., San Diego, CA (Berger); Sambrook *et al.*, MOLECULAR CLONING - A LABORATORY MANUAL (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, F.M. Ausubel *et al.*, eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1998) ("Ausubel"). Examples of techniques sufficient to direct persons of skill through *in vitro* amplification methods, including the polymerase chain reaction (PCR) the ligase chain reaction (LCR), Q β -replicase amplification and other RNA polymerase mediated techniques (e.g., NASBA) are found in Berger, Sambrook, and Ausubel, as well as Mullis *et al.*, U.S. Patent No. 4,683,202 (1987); PCR PROTOCOLS A GUIDE TO METHODS AND APPLICATIONS (Innis *et al.* eds), Academic Press, Inc., San Diego, CA (1990) (Innis); Arnheim & Levinson (October 1, 1990) *C&EN* 36-47; *The Journal Of NIH Research* 3:81-94 (1991); (Kwoh *et al.*, *Proc. Natl. Acad. Sci. USA* 86:1173 (1989); Guatelli *et al.*, *Proc. Natl. Acad. Sci. USA* 87:1874 (1990); Lomell *et al.*, *J. Clin. Chem* 35:1826 (1989); Landegren *et al.*, *Science* 241:1077-1080 (1988); Van Brunt, *Biotechnology* 8:291-294 (1990); Wu and Wallace, *Gene* 4:560 (1989); Barringer *et al.*, *Gene* 89:117 (1990); and Sooknanan and Malek, *Biotechnology* 13:563-564 (1995). Improved methods of cloning *in vitro* amplified nucleic acids are described in Wallace *et al.*, U.S. Pat. No. 5,426,039. Improved methods of amplifying large nucleic acids by PCR are summarized in Cheng *et al.*, *Nature* 369:684-685 (1994) and the references cited therein, in which PCR amplicons of up to 40kb are generated. One of skill will appreciate that essentially any RNA can be converted into a double stranded DNA suitable for restriction digestion, PCR expansion and sequencing using reverse transcriptase and a polymerase. See, Ausubel, Sambrook and Berger, *all supra*.

In another aspect, the present invention provides a method of increasing monooxygenase activity in a cell. The method includes performing whole genome shuffling of a plurality of genomic nucleic acids in the cell and selecting for one or more monooxygenase activity. In this aspect of the invention, the genomic nucleic acids can be from substantially any source. In a preferred embodiment of this aspect of the invention, the genomic nucleic acids are from a species or strain different from the cell. In a further preferred embodiment, the cell is of prokaryotic or eukaryotic origin.

Substantially any monooxygenase property can be selected for using the methods of the invention. A preferred property is the activity of the polypeptide towards a particular class of substrates. In preferred embodiment, the monooxygenase property is its ability to effect alkene epoxidation, alkane oxidation (e.g., hydroxylation, conversion to carboxylic acid, etc.), aromatic hydroxylation, N-dealkylation of alkylamines, S-dealkylation of reduced thio-organics, O-Dealkylation of alkyl ethers, oxidation of aryloxy phenols, conversion of aldehydes to acids, dehydrogenation, decarbonylation, oxidative dehalogenation of haloaromatics and halohydrocarbons, Baeyer-Villiger monooxygenation, modification of cyclosporins, hydroxylation of mevastatin, hydroxylation of fatty acids, hydroxylation/epoxidation of terpenes, conversion of cholesterol to pregnenolone, or oxygenation of sulfonylureas.

In a third aspect, the invention provides a DNA shuffling mixture comprising: at least three homologous DNAs, each of which is derived from a nucleic acid encoding a polypeptide or polypeptide fragment which encodes monooxygenase activity. In a preferred embodiment of this aspect of the invention, the at least three homologous DNAs are present in cell culture or *in vitro*.

Oligonucleotides for use as probes, e.g., in *in vitro* amplification methods, for use as gene probes, or as shuffling targets (e.g., synthetic genes or gene segments) are typically synthesized chemically according to the solid phase phosphoramidite triester method described by Beaucage and Caruthers, *Tetrahedron Letts.* 22(20):1859-1862, (1981) e.g., using an automated synthesizer, as described in Needham-VanDevanter *et al.*, *Nucleic Acids Res.*, 12:6159-6168 (1984). Oligonucleotides can also be custom made and ordered from a variety of commercial sources known to persons of skill.

A. Formats for Sequence Recombination

The methods of the invention entail performing recombination ("shuffling") and screening or selection to "evolve" individual genes, whole plasmids or viruses, multigene clusters, or even whole genomes (Stemmer, *Bio/Technology* 13:549-553 (1995)). Reiterative cycles of recombination and screening/selection can be performed to further evolve the nucleic acids of interest. Such techniques do not require the extensive analysis and computation required by conventional methods for polypeptide engineering. Shuffling allows the recombination of large numbers of mutations in a minimum number of selection cycles, in contrast to natural pair-wise recombination events (*e.g.*, as occur during sexual replication). Thus, the sequence recombination techniques described herein provide particular advantages in that they provide recombination between mutations in any or all of these, thereby providing a very fast way of exploring the manner in which different combinations of mutations can affect a desired result. In some instances, however, structural and/or functional information is available which, although not required for sequence recombination, provides opportunities for modification of the technique.

Sequence recombination can be achieved in many different formats and permutations of formats. Exemplary formats and examples for sequence recombination, referred to, *e.g.*, as "DNA shuffling," "fast forced evolution," or "molecular breeding," have been described in the following patents and patent applications: US Patent Application Serial No. 08/198,431, filed February 17, 1994, US Patent No. 5,605,793; PCT Application WO 95/22625 (Serial No. PCT/US95/02126), filed February 17, 1995; US Serial No. 08/425,684, filed April 18, 1995; Serial No. 08/537,874, filed October 30, 1995, Serial No. 08/564,955, filed November 30, 1995, Serial No. 08/621,859, filed March 25, 1996, US Serial No. 08/621,430, filed March 25, 1996; Serial No. PCT/US96/05480, filed April 18, 1996, Serial No. 08/650,400, filed May 20, 1996, Serial No. PCT/US97/17300, filed September 26, 1997, Serial No. PCT/US97/24239, filed December 17, 1997; Serial No. 98/354,922, filed July 15, 1999, Serial No. PCT/US98/05956, filed March 25, 1998; PCT Application WO 97/20078 (Serial No. PCT/US96/05480), filed April 18, 1996; PCT Application WO 97/35966, filed March 20, 1997; US Serial No. 08/675,502, filed July 3, 1996; US Serial No. 08/721,824, filed September 27, 1996; PCT Application WO 98/13487, filed September 26, 1997; "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination" Attorney Docket No. 018097-020720US filed July 15, 1998 by del Cardayre *et al.* (USSN

09/161,188); Stemmer, *Science* **270**:1510 (1995); Stemmer *et al.*, *Gene* **164**:49-53 (1995); Stemmer, *Bio/Technology* **13**:549-553 (1995); Stemmer, *Proc. Natl. Acad. Sci. U.S.A.* **91**:10747-10751 (1994); Stemmer, *Nature* **370**:389-391 (1994); Crameri *et al.*, *Nature Medicine* **2**(1):1-3 (1996); Crameri *et al.*, *Nature Biotechnology* **14**:315-319 (1996), and
5 PCT Application WO 98/42832 (Serial No. PCT/US98/05956), filed March 25, 1998, each of which is incorporated by reference in its entirety for all purposes.

Gene shuffling and family shuffling provide two of the most powerful methods available for improving and "migrating" (gradually changing the type of reaction, substrate or activity of a selected enzyme) the functions of biocatalysts. In family shuffling,
10 homologous sequences, *e.g.*, from different species or chromosomal positions, are recombined. In gene shuffling, a single sequence is mutated or otherwise altered and then recombined. These formats share some common principles.

The breeding procedure starts with at least two substrates that generally show substantial sequence identity to each other (*i.e.*, at least about 30%, 50%, 70%, 80% or 90%
15 sequence identity), but differ from each other at certain positions. The difference can be any type of mutation, for example, substitutions, insertions and deletions. Often, different segments differ from each other in about 5-20 positions. For recombination to generate increased diversity relative to the starting materials, the starting materials must differ from each other in at least two nucleotide positions. That is, if there are only two substrates, there
20 should be at least two divergent positions. If there are three substrates, for example, one substrate can differ from the second at a single position, and the second can differ from the third at a different single position. The starting DNA segments can be natural variants of each other, for example, allelic or species variants. The segments can also be from nonallelic genes showing some degree of structural and usually functional relatedness (*e.g.*,
25 different genes within a superfamily, such as the arene monooxygenase super family). The starting DNA segments can also be induced variants of each other. For example, one DNA segment can be produced by error-prone PCR replication of the other, or by substitution of a mutagenic cassette. Induced mutants can also be prepared by propagating one (or both) of the segments in a mutagenic strain. In these situations, strictly speaking, the second DNA
30 segment is not a single segment but a large family of related segments. The different segments forming the starting materials are often the same length or substantially the same length. However, this need not be the case; for example; one segment can be a subsequence

of another. The segments can be present as part of larger molecules, such as vectors, or can be in isolated form.

The starting DNA segments are recombined by any of the sequence recombination formats provided herein to generate a diverse library of recombinant DNA segments. Such a library can vary widely in size from having fewer than 10 to more than 10^5 , 10^9 , 10^{12} or more members. In some embodiments, the starting segments and the recombinant libraries generated will include full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. In other embodiments, the recombinant DNA segments in the library can be inserted into a common vector providing sequences necessary for expression before performing screening/selection.

1. Use of Restriction Enzyme Sites to Recombine Mutations

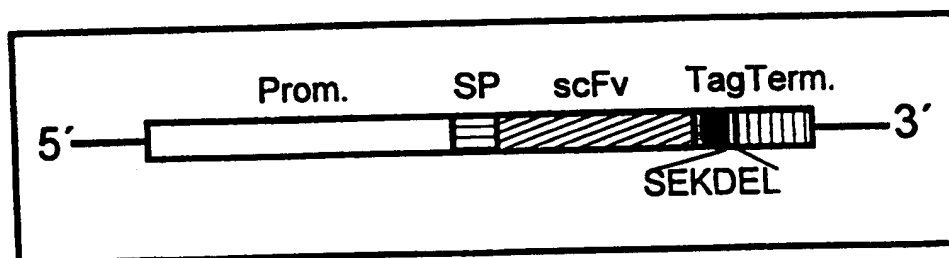
In some situations it is advantageous to use restriction enzyme sites in nucleic acids to direct the recombination of mutations in a nucleic acid sequence of interest. These techniques are particularly preferred in the evolution of fragments that cannot readily be shuffled by existing methods due to the presence of repeated DNA or other problematic primary sequence motifs. These situations also include recombination formats in which it is preferred to retain certain sequences unmutated. The use of restriction enzyme sites is also preferred for shuffling large fragments (typically greater than 10 kb), such as gene clusters that cannot be readily shuffled and "PCR-amplified" because of their size. Although fragments up to 50 kb have been reported to be amplified by PCR (Barnes, *Proc. Natl. Acad. Sci. U.S.A.* 91:2216-2220 (1994)), it can be problematic for fragments over 10 kb, and thus alternative methods for shuffling in the range of 10 - 50 kb and beyond are preferred. Preferably, the restriction endonucleases used are of the Class II type (Sambrook, Ausubel and Berger, *supra*) and of these, preferably those which generate nonpalindromic sticky end overhangs such as AlwI, SfiI or BstXI. These enzymes generate nonpalindromic ends that allow for efficient ordered reassembly with DNA ligase. Typically, restriction enzyme (or endonuclease) sites are identified by conventional restriction enzyme mapping techniques (Sambrook, Ausubel, and Berger, *supra.*), by analysis of sequence information for that gene, or by introduction of desired restriction sites into a nucleic acid sequence by synthesis (*i.e.* by incorporation of silent mutations).



PCT
WELTORGANISATION FÜR GEISTIGES EIGENTUM
Internationales Büro
INTERNATIONALE ANMELDUNG VERÖFFENTLICHT NACH DEM VERTRAG ÜBER DIE
INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT)

(51) Internationale Patentklassifikation ⁶ : C12N 15/82, C07K 16/44, C12N 15/13, A01H 5/00, A01N 63/00		A1	(11) Internationale Veröffentlichungsnummer: WO 98/42852
		(43) Internationales Veröffentlichungsdatum:	1. Oktober 1998 (01.10.98)
(21) Internationales Aktenzeichen: PCT/EP98/01731		(81) Bestimmungsstaaten: AL, AU, BG, BR, BY, CA, CN, CZ, GE, HU, ID, IL, JP, KR, KZ, LT, LV, MX, NO, NZ, PL, RO, RU, SG, SI, SK, TR, UA, US, UZ, VN, eurasisches Patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europäisches Patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).	
(22) Internationales Anmeldedatum: 24. März 1998 (24.03.98)			
(30) Prioritätsdaten: 197 12 507.7 25. März 1997 (25.03.97) DE			
(71) Anmelder (für alle Bestimmungsstaaten ausser US): BASF AK- TIENGESELLSCHAFT [DE/DE]; D-67056 Ludwigshafen (DE).		Veröffentlicht <i>Mit internationalem Recherchenbericht. Vor Ablauf der für Änderungen der Ansprüche zugelassenen Frist; Veröffentlichung wird wiederholt falls Änderungen eintreffen.</i>	
(72) Erfinder; und (75) Erfinder/Anmelder (nur für US): LERCHL, Jens [DE/DE]; Im Steg 36, D-68526 Ladenburg (DE). MÖLLER, Achim [DE/DE]; Im Zaunrücken 10, D-67269 Grünstadt (DE). SCHMIDT, Ralf-Michael [DE/DE]; Am Schlossgarten 9 d, D-67489 Kirrweiler (DE). SCHIFFER, Helmut [DE/DE]; Theodor-Heuss-Strasse 31, D-67112 Mutterstadt (DE). RABE, Udo [DE/DE]; Wachenheimer Strasse 2, D-67125 Dannstadt-Schauernheim (DE). CONRAD, Udo [DE/DE]; Corrensstrasse 3, D-06466 Gatersleben (DE).			
(74) Gemeinsamer Vertreter: BASF AKTIENGESELLSCHAFT; D-67056 Ludwigshafen (DE).			

(54) Title: **EXPRESSION OF HERBICIDE-BINDING POLYPEPTIDES IN PLANTS TO PRODUCE HERBICIDE TOLERANCE**
(54) Bezeichnung: **EXPRESSION VON HERBIZID-BINDENDEN POLYPEPTIDEN IN PFLANZEN ZUR ERZEUGUNG VON
HERBIZIDTOLERANZ**



☐ USP-Promoter

☐ LeB4-Signalpeptid

☐ c-myc-Tag

☐ CaMV 35 Terminator

☐ single chain Fv

(57) Abstract

Disclosed is a method for producing herbicide-tolerant plants by expression of a herbicide-binding antibody in plants.

(57) Zusammenfassung

Verfahren zur Herstellung von Herbizid-toleranten Pflanzen durch Expression eines Herbizid-bindenden Antikörpers in den Pflanzen.

LEDIGLICH ZUR INFORMATION

Codes zur Identifizierung von PCT-Vertragsstaaten auf den Kopfbögen der Schriften, die internationale Anmeldungen gemäss dem PCT veröffentlichen.

AL	Albanien	ES	Spanien	LS	Lesotho	SI	Slowenien
AM	Armenien	FI	Finnland	LT	Litauen	SK	Slowakei
AT	Österreich	FR	Frankreich	LU	Luxemburg	SN	Senegal
AU	Australien	GA	Gabun	LV	Lettland	SZ	Swasiland
AZ	Aserbaidschan	GB	Vereinigtes Königreich	MC	Monaco	TD	Tschad
BA	Bosnien-Herzegowina	GE	Georgien	MD	Republik Moldau	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagaskar	TJ	Tadschikistan
BE	Belgien	GN	Guinea	MK	Die ehemalige jugoslawische Republik Mazedonien	TM	Turkmenistan
BF	Burkina Faso	GR	Griechenland	ML	Mali	TR	Türkei
BG	Bulgarien	HU	Ungarn	MN	Mongolei	TT	Trinidad und Tobago
BJ	Benin	IE	Irland	MR	Mauretanien	UA	Ukraine
BR	Brasilien	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Island	MX	Mexiko	US	Vereinigte Staaten von Amerika
CA	Kanada	IT	Italien	NE	Niger	UZ	Usbekistan
CF	Zentralafrikanische Republik	JP	Japan	NL	Niederlande	VN	Vietnam
CG	Kongo	KE	Kenia	NO	Norwegen	YU	Jugoslawien
CH	Schweiz	KG	Kirgisistan	NZ	Neuseeland	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Demokratische Volksrepublik Korea	PL	Polen		
CM	Kamerun	KR	Republik Korea	PT	Portugal		
CN	China	KZ	Kasachstan	RO	Rumänien		
CU	Kuba	LC	St. Lucia	RU	Russische Föderation		
CZ	Tschechische Republik	LI	Liechtenstein	SD	Sudan		
DE	Deutschland	LK	Sri Lanka	SE	Schweden		
DK	Dänemark	LR	Liberia	SG	Singapur		
EE	Estland						

Expression von Herbizid-bindenden Polypeptiden in Pflanzen zur
Erzeugung von Herbizidtoleranz

5 Beschreibung

Die vorliegende Erfindung betrifft ein Verfahren zur Herstellung von herbizidtoleranten Pflanzen durch Expression eines exogenen Herbizid-bindenden Polypeptides in Pflanzen oder Pflanzenteilen.

- 10 Die Erfindung betrifft weiterhin die Verwendung der entsprechenden Nukleinsäuren codierend für ein Polypeptid, einen Antikörper oder Teilen eines Antikörpers mit Herbizid-bindenden Eigenschaften in transgenen Pflanzen und die auf diese Weise transformierte Pflanze selbst.

15

Es ist bekannt, daß mit Hilfe von gentechnischen Verfahren gezielt Fremdgene in das Genom einer Pflanze übertragen werden können. Dieser Prozeß wird als Transformation und die resultierenden Pflanzen werden als transgene Pflanzen bezeichnet. Transgene

- 20 Pflanzen werden derzeit in unterschiedlichen biotechnologischen Bereichen eingesetzt. Beispiele sind insektenresistente Pflanzen (Vaek et al. Plant Cell 5 (1987), 159-169), virusresistente Pflanzen (Powell et al. Science 232 (1986), 738-743) und ozonresistente Pflanzen (Van Camp et al. BioTech. 12 (1994), 165-168).
25 Beispiele für gentechnisch erzielte Qualitätssteigerungen sind: Erhöhung der Haltbarkeit von Früchten (Oeller et al. Science 254 (1991), 437-439), Erhöhung der Stärkeproduktion in Kartoffelknollen (Stark et al. Science 242 (1992), 419), Veränderung der Stärke- (Visser et al. Mol. Gen. Genet. 225 (1991), 289-296) und
30 Lipidzusammensetzung (Voelker et al. Science 257 (1992), 72-74) und Produktion pflanzenfremder Polymere (Poirer et al. Science 256 (1992), 520-523).

- Ein wichtiges Ziel der pflanzenmolekulargenetischen Arbeiten ist
35 die Erzeugung von Herbizidtoleranz. Die Herbizidtoleranz ist gekennzeichnet durch eine in Art oder Höhe gesteigerten Verträglichkeit der Pflanze oder von Pflanzenteilen gegenüber dem applizierten Herbizid. Diese kann auf verschiedene Arten bewerkstelligt werden. Die bekannten Methoden sind die Nutzung eines Metabolismusgens wie z.B. das pat-Gen in Zusammenhang mit der Glufosinat-Resistenz (WO 8705629) oder einem gegenüber dem Herbizid
40 resistenten Zielenzym wie im Falle der Enolpyruvylshikimat-3-Phosphat-Synthase (WO 9204449), die resistent ist gegen Glyphosat, sowie die Verwendung eines Herbizids in Zell- und Ge-
45 webekultur zur Selektion toleranter Pflanzenzellen und daraus re-

2

sultierender resistenter Pflanzen wie bei Acetyl CoA Carboxylase Hemmstoffen beschrieben (US 5162602, US 5290696).

Antikörper sind Proteine als Bestandteil des Immunsystems. Allen
5 Antikörpern gemeinsam ist ihre räumliche, globuläre Struktur, der
Aufbau aus leichter und schwerer Kette sowie ihre prinzipielle
Fähigkeit, Moleküle oder Teile einer Molekülstruktur mit hoher
Spezifität binden zu können (Alberts et al., in: Molekularbiolo-
gie der Zelle, 2. Auflage 1990, VCH Verlag, ISBN 3-527-27983-0,
10 1198-1237). Aufgrund dieser Eigenschaften wurden Antikörper für
vielfältige Aufgaben genutzt. Man unterscheidet dabei die Anwen-
dung der Antikörper im tierischen und menschlichen Organismen,
die sie produzieren, die sogenannte in-situ Anwendungen und die
ex-situ Anwendungen, d.h. die Nutzung der Antikörper nach Isola-
15 tion aus den produzierenden Zellen oder Organismen (Whitelam und
Cockburn, TIPS Vol.1 , 8 (1996), 268-272).

Die Verwendung hybrider somatischer Zelllinien (Hybridomas) als
Quelle für Antikörper gegen ganz bestimmte Antigene geht auf Ar-
20 beiten von Köhler und Milstein zurück (Nature 256 (1975) 495-97).
Nach diesem Verfahren lassen sich sogenannte monoklonale Antikör-
per herstellen, die eine einheitliche Struktur besitzen und durch
Zellfusion erzeugt werden. Dabei werden Milzzellen einer immuni-
sierten Maus mit Zellen eines Mausmyeloms fusioniert. So entste-
25 hen Hybridomazellen, die sich unbegrenzt vermehren. Gleichzeitig
sezernieren die Zellen spezifische Antikörper gegen das Antigen,
mit dem die Maus immunisiert worden war. Die Milzzellen liefern
die Fähigkeit zur Antikörperproduktion, während die Myelomzellen
die unbegrenzte Wachstumsfähigkeit und die kontinuierliche Anti-
30 körpersekretion beisteuern. Da jede Hybridomazelle sich als Klon
von einer einzigen B-Zelle ableitet, besitzen alle erzeugten An-
tikörpermoleküle dieselbe Struktur einschließlich der Antigenbin-
dungsstelle. Diese Methode hat die Anwendung von Antikörpern
stark gefördert, da jetzt Antikörper mit einer einzigen, bekann-
35 ten Spezifität und einer homogenen Struktur unbegrenzt zur Verfü-
gung stehen. Monoklonale Antikörper finden breite Anwendung in
der Immundiagnostik und als Therapeutika.

Seit einigen Jahren gibt es die sogenannte Phagen-Display-Methode
40 zur Herstellung von Antikörpern, bei der das Immunsystem und die
verschiedenen Immunisierungen im Tier umgangen werden. Hierbei
wird die Affinität und Spezifität des Antikörpers in vitro maßge-
schneidert (Winter et al., Ann. Rev. Immunol. 12 (1994), 433-455;
Hoogenboom TIBTech Vol 15 (1997), 62 -70). Gensegmente, die die
45 kodierende Sequenz der variablen Region von Antikörpern enthält,
d.h. die Antigen-Bindestelle, werden mit Genen für das Hüllpro-
tein eines Bakteriophagen fusioniert. Dann infiziert man Bakte-

3

rien mit Phagen, die solche Fusionsgene enthalten. Die entstehenden Phagenpartikel besitzen nun Hüllen mit dem antikörperähnlichen Fusionsprotein, wobei die antikörperbindende Domäne nach außen zeigt. Aus einer solchen Phagen-Display-Bibliothek läßt sich nun der Phage isolieren, der das gewünschte Antikörperfragment enthält und spezifisch an ein bestimmtes Antigen bindet. Jeder so isolierte Phage erzeugt ein monoklonales, antigenbindendes Polypeptid, das einem monoklonalen Antikörper entspricht. Die Gene für die Antigenbindungsstelle, die für jeden Phagen einzigartig sind, kann man aus der Phagen-DNA isolieren und zur Konstruktion vollständiger Antikörpergene einsetzen.

Auf dem Gebiet des Pflanzenschutzes wurden Antikörper insbesondere als analytisches Mittel ex-situ zum qualitativen und quantitativen Nachweis von Antigenen genutzt. Dies schließt den Nachweis von Pflanzeninhaltsstoffen, Herbiziden oder Fungiziden in Trinkwasser (Sharp et al. (1991) ACS Symp Ser., 446 (Pestic. Residues Food Saf.) 87-95), Bodenproben (WO 9423018) oder in Pflanzen oder Pflanzenteilen sowie die Nutzung von Antikörpern als Hilfsmittel zur Reinigung von gebundenen Molekülen ein.

Die Produktion von Immunglobulinen in Pflanzen wurde erstmals von Hiatt et al., Nature, 342 (1989), 76 - 78 beschrieben. Das Spektrum reicht von Ein-Ketten-Antikörpern bis zu multimeren sekretorischen Antikörpern (J. Ma und Mich Hein, 1996, Annuals New York Academy of Sciences, 72 - 81).

Neuere Versuche nutzen Antikörper in-situ zur Pathogenabwehr in Pflanzen, insbesondere von Viruserkrankungen durch Expression von spezifischen Antikörpern oder Teilen davon gerichtet gegen Virus-hüllproteine in Pflanzenzellen (Tavladoraki et al., Nature 366 (1993), 469-472; Voss et al., Mol. Breeding 1 (1995), 39-50).

Ein analoger Ansatz ist auch zur Abwehr der Infektion der Pflanze durch Nematoden genutzt worden (Rosso et al., Biochem Biophys Res Com, 220 (1996) 255-263). Für eine pharmazeutische Anwendung sind Beispiele bekannt, die die Antikörper-Expression in-situ in Pflanzen für eine orale Immunisierung nutzen (Ma et al., Science 268 (1995), 716-719; Mason und Arntzen, Tibtech Vol 13 (1996), 388-392). Von der Pflanze gebildete Antikörper werden dabei aus Pflanzen oder für den Verzehr geeigneten Pflanzenteilen über den Mund, Rachen oder Verdauungstrakt dem Körper zugeführt und verursachen einen wirksamen Immunschutz. Weiterhin wurde in Pflanzen bereits ein Ein-Ketten-Antikörper (single chain antibody) gegen das niedermolekulare Pflanzenhormon Abscisinsäure exprimiert und eine verringerte Pflanzenhormonverfügbarkeit aufgrund von Absci-

4

sinsäurebindung in der Pflanze beobachtet (Artsaenko et al., The Plant Journal (1995) 8 (5), 745-750).

Die chemische Unkrautbekämpfung in agrarwirtschaftlich bedeuten-
5 den Kulturen setzt den Einsatz von hochselektiven Herbiziden vor-
aus. In einigen Fällen ist es jedoch schwierig, Herbizide mit
ausreichender Selektivität zu entwickeln, die keine Schädigung
der Ertragspflanze verursachen. Die Einführung von Herbizid-resi-
stenten oder -toleranten Kulturpflanzen kann zur Lösung dieses
10 Problems beitragen.

Der Entwicklung von Herbizid-resistenten Kulturpflanzen durch Ge-
webekultur oder Samenmutagenese und natürliche Auswahl sind Gren-
zen gesetzt. So können nur diejenigen Pflanzen über Gewebekultur-
15 techniken manipuliert werden, deren Regeneration zu ganzen Pflan-
zen aus Zellkulturen gelingt. Außerdem können Kulturpflanzen nach
Mutagenese und Selektion unerwünschte Eigenschaften zeigen, die
durch teilweise mehrmalige Rückkreuzungen wieder beseitigt werden
müssen. Auch wäre die Einbringung einer Resistenz durch Kreuzung
20 auf Pflanzen der selben Art beschränkt.

Aus diesen Gründen ist der gentechnische Ansatz, ein für die
Resistenz codierendes Gen zu isolieren und in Kulturpflanzen ge-
zielt zu übertragen, dem klassischen Züchtungsverfahren überle-
25 gen.

Die molekularbiologische Entwicklung von Herbizid-toleranten bzw.
Herbizid-resistenten Kulturpflanzen setzt bisher voraus, daß der
Wirkmechanismus des Herbizides in der Pflanze bekannt ist und daß
30 Gene, die Resistenz gegen das Herbizid vermitteln gefunden werden
können. Viele gegenwärtig kommerziell genutzten Herbizide wirken,
indem sie ein Enzym einer essentiellen Aminosäure-, Lipid- oder
Pigmentbiosynthese blockieren. Durch Veränderung der Gene dieser
Enzyme dergestalt, daß das Herbizid nicht mehr gebunden werden
35 kann und durch Einbringung dieser veränderten Gene in Kultur-
pflanzen läßt sich Herbizid-Toleranz erzeugen. Alternativ können
zum Beispiel in der Natur analoge Enzyme beispielsweise in Mikro-
organismen gefunden werden, die eine natürliche Resistenz gegen-
über dem Herbizid zeigen. Dieses Resistenz vermittelnde Gen wird
40 aus einem derartigen Mikroorganismus isoliert, in geeignete Vek-
toren umklontiert und anschließend nach erfolgreicher Transforma-
tion in Herbizid-sensitiven Kulturpflanzen zur Expression ge-
bracht (WO 96/38567).

5

Aufgabe der vorliegenden Erfindung war die Entwicklung eines neuartigen allgemein einsetzbaren, gentechnologischen Verfahrens zur Erzeugung von Herbizid-toleranten transgenen Pflanzen.

- 5 Diese Aufgabe wurde überraschenderweise gelöst durch ein Verfahren der Expression eines exogenen Polypeptides, Antikörpers oder Teilen eines Antikörpers mit Herbizid-bindenden Eigenschaften in den Pflanzen.
- 10 Ein erster Gegenstand der vorliegenden Erfindung betrifft die Herstellung eines Herbizid-bindenden Antikörpers und die Klonierung des zugehörigen Gens bzw. Genfragmentes.

Es wird zunächst ein geeigneter Antikörper erzeugt, der das

15 Herbizid bindet. Dies kann u.a. durch Immunisierung eines Wirbeltiers, meist Maus, Ratte, Hund, Pferd, Esel oder Ziege mit einem Antigen erfolgen. Das Antigen ist dabei eine herbizid wirksame Verbindung, die über eine funktionelle Gruppe an einen höher-molekularen Träger wie Rinderserumalbumin (BSA), Hühnereiweiß

20 (Ovalbumin) keyhole limpet hemocyanin (KLH) oder andere Träger gekoppelt oder assoziiert vorliegt. Die Immunantwort wird nach mehrmaliger Antigenapplikation mit gängigen Methoden nachvollzogen und so ein geeignetes Antiserum isoliert. Dieser Ansatz liefert zunächst ein polyklonales Serum, das Antikörper mit unterschiedlichen Spezifitäten enthält. Für den gezielten in-situ Gebrauch ist es notwendig, die für einen einzelnen spezifischen

25 monoklonalen Antikörper codierende Gensequenz zu isolieren. Zu diesem Zweck stehen verschiedene Wege offen. Der erste Ansatz nutzt die Fusion von Antikörper-produzierenden Zellen mit Krebszellen zu einer ständig Antikörper produzierenden Hybridomazellkultur, die durch Vereinzelung der enthaltenen Klone letztlich zu einer homogenen, einen definierten monoklonalen Antikörper produzierenden Zelllinie führt.

- 35 Aus einer derartigen monoklonalen Zelllinie wird die cDNA für den Antikörper bzw. Teile des Antikörpers, den sog. Ein-Ketten-Antikörper (single chain antibody - scFv) isoliert. Diese cDNA-Sequenzen können dann in Expressionskassetten kloniert und zur funktionellen Expression in prokaryotischen und eukaryotischen
- 40 Organismen, einschließlich Pflanzen genutzt werden.

Es ist auch möglich über Phagen-Display-Banken Antikörper zu selektieren, die Herbizidmoleküle binden und katalytisch in ein Produkt mit nicht herbiziden Eigenschaften umsetzen. Methoden zur

45 Herstellung katalytischer Antikörper sind in Janda et al., Science 275 (1997) 945-948, Chemical selection for catalysis in combinatorial Antibody libraries; Catalytic Antibodies, 1991,

6

Ciba Foundation Symposium 159, Wiley- Interscience Publication beschrieben. Durch Klonierung des Gens dieses katalytischen Antikörpers und dessen Expression in einer Pflanze kann im Prinzip ebenfalls eine Herbizid-resistente Pflanze erzeugt werden.

5

- Gegenstand der Erfindung sind insbesondere Expressionskassetten, deren kodierende Sequenz für ein Herbizid-bindendes Polypeptid oder dessen funktionelles Äquivalent codiert, sowie deren Verwendung zur Herstellung einer Herbizid-toleranten Pflanze. Die
- 10 Nukleinsäuresequenz kann dabei z.B. eine DNA- oder eine cDNA-Sequenz sein. Zur Insertion in eine erfindungsgemäße Expressionskassette geeignete kodierende Sequenzen sind beispielsweise solche, die eine DNA-Sequenz aus einer Hybridomazelle enthalten, die für ein Polypeptid mit Herbizid-bindenden Eigenschaften codiert
- 15 und die dem Wirt Resistenz gegen Inhibitoren pflanzlicher Enzyme verleihen.

- Die erfindungsgemäßen Expressionskassetten beinhalten außerdem regulative Nukleinsäuresequenzen, welche die Expression der co-
- 20 dierenden Sequenz in der Wirtszelle steuern. Gemäß einer bevorzugten Ausführungsform umfaßt eine erfindungsgemäße Expressionskassette stromaufwärts, d.h. am 5'-Ende der codierenden Sequenz einen Promotor und stromabwärts, d.h. am 3'-Ende ein Polyadenylierungssignal und gegebenenfalls weitere regulatorische Ele-
- 25 mente, welche mit der dazwischenliegenden codierenden Sequenz für das Polypeptid mit Herbizid-bindenden Eigenschaften und/oder Transitpeptid operativ verknüpft sind. Unter einer operativen Verknüpfung versteht man die sequenzielle Anordnung von Promotor, codierender Sequenz, Terminator und ggf. weiterer regulativer
- 30 Elemente derart, daß jedes der regulativen Elemente seine Funktion bei der Expression der codierenden Sequenz bestimmungsgemäß erfüllen kann. Die zur operativen Verknüpfung bevorzugten aber nicht darauf beschränkten Sequenzen sind Targeting-Sequenzen zur Gewährleistung der subzellulären Lokalisation im Apoplasten, in
- 35 der Plasmamembran, in der Vakuole, in Plastiden, ins Mitochondrium, im Endoplasmatischen Retikulum (ER), im Zellkern, in Ölkörperchen oder anderen Kompartimenten und Translationsverstärker wie die 5'-Führungssequenz aus dem Tabak Mosaic Virus (Gallie et al., Nucl. Acids Res. 15 (1987) 8693-8711).

40

- Als Promotoren der erfindungsgemäßen Expressionskassette ist grundsätzlich jeder Promotor geeignet, der die Expression von Fremdgenen steuern kann. Vorzugsweise verwendet man insbesondere einen pflanzlichen Promotor oder einen Promotor, der einem Pflanzenvirus entstammt. Insbesondere bevorzugt ist der CaMV 35S-Promotor aus dem Blumenkohl-Mosaik-Virus (Franck et al., Cell
- 45 21(1980) 285-294). Dieser Promotor enthält unterschiedliche Er-

7

kennungssequenzen für transkriptionale Effektoren, die in ihrer Gesamtheit zu einer permanenten und konstitutiven Expression des eingeführten Gens führen (Benfey et al., EMBO J. 8 (1989) 2195-2202).

5

Die erfindungsgemäße Expressionskassette kann auch einen chemisch induzierbaren Promotor enthalten, durch den die Expression des exogenen Polypeptids in der Pflanze zu einem bestimmten Zeitpunkt gesteuert werden kann. Derartige Promotoren wie z.B. der PRP1-Promotor (Ward et al., Plant.Mol.Biol.22(1993), 361-366), ein durch Salizylsäure induzierbarer Promotor (WO 95/1919443), ein durch Benzenesulfonamid-induzierbarer (EP 388186), ein durch Abscisinsäure-induzierbarer (EP335528) bzw. ein durch Ethanol- oder Cyclohexanon-induzierbarer (WO9321334) Promotor sind der Literatur beschrieben und können u.a. verwendet werden.

Weiterhin sind insbesondere solche Promotoren bevorzugt, die Expression in Geweben oder Pflanzenteilen sicherstellen, in denen sich die Herbizidwirkung entfaltet. Insbesondere zu nennen sind Promotoren, die eine Blatt-spezifische Expression gewährleisten. Zu nennen sind der Promotor der cytosolischen FBPase aus Kartoffel oder der ST-LSI Promotor aus Kartoffel (Stockhaus et al., EMBO J. 8 (1989) 2445-245).

Mit Hilfe eines samenspezifischen Promotors konnten Einketten-Antikörper stabil bis zu 0,67% des gesamten löslichen Samenproteins in den Samen transgener Tabakpflanzen exprimiert werden (Fiedler und Conrad, Bio/Technology 10(1995), 1090-1094). Da auch eine Expression in ausgesäten oder keimenden Samen möglich und im Sinne der vorliegenden Erfindung erwünscht sein kann, sind entsprechend Keimungs- und Samen-spezifische Promotoren ebenfalls erfindungsgemäß bevorzugte regulative Elemente. Die erfindungsgemäße Expressionskassette kann daher beispielsweise einen samenspezifischen Promotor (bevorzugt den USP- oder LEB4-Promotor, das LEB4-Signalpeptid, das zu exprimierende Gen und ein ER-Retentionssignal enthalten. Der Aufbau der Kassette ist in der Abbildung 1 am Beispiel eines Einketten-Antikörpers (scFv-Gen) schematisch beispielhaft dargestellt.

Die Herstellung einer erfindungsgemäßen Expressionskassette erfolgt durch Fusion eines geeigneten Promotors mit einer geeigneten Polypeptid-DNA und vorzugsweise einer zwischen Promotor und Polypeptid-DNA insertierten für ein Chloroplasten-spezifisches Transitpeptid kodierenden DNA sowie einem Polyadenylierungssignal nach gängigen Rekombinations- und Klonierungstechniken, wie sie beispielsweise in T. Maniatis, E.F. Fritsch und J. Sambrook, Molecular Cloning: A Laboratory manual, Cold Spring Harbor Labora-

tory, Cold Spring Harbor, NY (1989) sowie in T.J. Silhavy, M.L. Berman und L.W. Enquist, Experiments with Gene Fusions, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1984) und in Ausubel, F.M. et al., Current Protocols in Molecular Biology, Greene Publishing Assoc. and Wiley-Interscience (1987) beschrieben sind.

Insbesondere bevorzugt sind Sequenzen, die ein Targeting in den Apoplasten, Plastiden, die Vakuole, in die Plasmamembran, das Mitochondrium, das Endoplasmatische Retikulum (ER) oder durch ein Fehlen entsprechender operativer Sequenzen einen Verbleib im Kompartiment des Entstehens, dem Zytosol, gewährleisten (Kermode, Crit. Rev. Plant Sci. 15, 4 (1996), 285-293). Für die Menge der Proteinakkumulation in transgenen Pflanzen besonders förderlich erwiesen hat sich eine Lokalisation im ER sowie der Zellwand (Schouten et al., Plant Mol. Biol. 30 (1996), 781-792; Artsaenko et al., Plant J. 8 (1995) 745-750).

Gegenstand der Erfindung sind auch Expressionskassetten, deren kodierende Sequenz für ein Herbizid-bindendes Fusionsprotein kodiert, wobei Teil des Fusionsproteins ein Transitpeptid ist, das die Translokation des Polypeptides steuert. Besonders bevorzugt sind Chloroplasten-spezifische Transitpeptide, welche nach Translokation des Herbizid-bindenden Polypeptides in die Pflanzenchloroplasten vom Herbizid-bindenden Polypeptid-Teil enzymatisch abgespalten werden. Insbesondere bevorzugt ist das Transitpeptid abgeleitet von plastidärer Transketolase (TK) oder einem funktionellen Äquivalent dieses Transitpeptids (z.B. das Transitpeptid der kleinen Untereinheit der Rubisco oder der Ferredoxin NADP Oxidoreduktase).

Die zur Herstellung erfindungsgemäßer Expressionskassetten erforderliche Polypeptid-DNA oder -cDNA wird vorzugsweise mit Hilfe der Polymerase-Kettenreaktion (PCR) amplifiziert. Verfahren zur DNA-Amplifikation mittels PCR sind bekannt, beispielsweise aus Innis et al., PCR Protocols, A Guide to Methods and Applications, Academic Press (1990). Zweckmäßigerweise können die PCR-erzeugten DNA-Fragmente durch Sequenzanalyse zur Vermeidung von Polymerasefehlern in zu exprimierenden Konstrukten überprüft werden.

Die insertierte Nucleotid-Sequenz codierend für ein Herbizid-bindendes Polypeptid kann synthetisch hergestellt oder natürlich gewonnen sein oder eine Mischung aus synthetischen und natürlichen DNA-Bestandteilen enthalten. Im allgemeinen werden synthetische Nucleotid-Sequenzen mit Codons erzeugt, die von Pflanzen bevorzugt werden. Diese von Pflanzen bevorzugten Codons können aus Codons mit der höchsten Proteinhäufigkeit bestimmt werden, die in

den meisten interessanten Pflanzenspezies exprimiert werden. Bei der Präparation einer Expressionskassette können verschiedene DNA-Fragmente manipuliert werden, um eine Nucleotid-Sequenz zu erhalten, die zweckmäßigerweise in der korrekten Richtung liest und die mit einem korrekten Leseraster ausgestattet ist. Für die Verbindung der DNA-Fragmente miteinander können an die Fragmente Adaptoren oder Linker angesetzt werden.

Zweckmäßigerweise sollten die erfindungsgemäßen Promotor- und Terminator-Regionen in Transkriptionsrichtung mit einem Linker oder Polylinker, der eine oder mehrere Restriktionsstellen für die Insertion dieser Sequenz enthält, versehen werden. In der Regel hat der Linker 1 bis 10, meistens 1 bis 8, vorzugsweise 2 bis 6 Restriktionsstellen. Im allgemeinen hat der Linker innerhalb der regulatorischen Bereiche eine Größe von weniger als 100 bp, häufig weniger als 60 bp, mindestens jedoch 5 bp. Der erfindungsgemäße Promotor kann sowohl nativ bzw. homolog als auch fremdartig bzw. heterolog zur Wirtspflanze sein. Die erfindungsgemäße Expressionskassette beinhaltet in der 5'-3'-Transkriptionsrichtung den erfindungsgemäßen Promotor, eine beliebige Sequenz und eine Region für die transkriptionale Termination. Verschiedene Terminationsbereiche sind gegeneinander beliebig austauschbar.

Ferner können Manipulationen, die passende Restriktionsschnittstellen bereitstellen oder die überflüssige DNA oder Restriktionsschnittstellen entfernen, eingesetzt werden. Wo Insertionen, Deletionen oder Substitutionen wie z.B. Transitionen und Transversionen in Frage kommen, können in vitro-Mutagenese, "primer-repair", Restriktion oder Ligation verwendet werden. Bei geeigneten Manipulationen, wie z.B. Restriktion, "chewing-back" oder Auffüllen von Überhängen für "bluntends", können komplementäre Enden der Fragmente für die Ligation zur Verfügung gestellt werden.

Von besonderer Bedeutung für den erfindungsgemäßen Erfolg ist das Anhängen des spezifischen ER-Retentionssignals SEXDEL (Schuoten, A. et al. Plant Mol. Biol. 30 (1996), 781 - 792), die durchschnittliche Expressionshöhe wird damit verdreifacht bis vervierfacht. Es können auch andere Retentionssignale, die natürlicherweise bei im ER lokalisierten pflanzlichen und tierischen Proteinen vorkommen, für den Aufbau der Kassette eingesetzt werden.

Bevorzugte Polyadenylierungssignale sind pflanzliche Polyadenylierungssignale, vorzugsweise solche, die im wesentlichen T-DNA-Polyadenylierungssignale aus *Agrobacterium tumefaciens*, insbesondere des Gens 3 der T-DNA (Octopin Synthase) des Ti-Plasmids

pTiACH5 entsprechen (Gielen et al., EMBO J. 3 (1984) 835 ff) oder funktionelle Äquivalente.

Eine erfindungsgemäße Expressionskassette kann beispielsweise
5 einen konstitutiven Promotor (bevorzugt den CaMV 35 S-Promotor),
das LeB4-Signalpeptid, das zu exprimierende Gen und das ER-Retentionssignal enthalten. Der Aufbau der Kassette ist in Abbildung 2 am Beispiel eines Einketten-Antikörpers (scFv-Gen) schematisch dargestellt. Als ER-Retentionssignal wird bevorzugt die Aminosäuresequenz KDEL (Lysin, Asparaginsäure, Glutaminsäure, Leucin)
10 verwendet.

Vorzugsweise wird die fusionierte Expressionskassette, die für ein Polypeptid mit Herbizid-bindenden Eigenschaften codiert, in
15 einen Vektor, beispielsweise pBin19, kloniert, der geeignet ist, Agrobacterium tumefaciens zu transformieren. Mit einem solchen Vektor transformierte Agrobakterien können dann in bekannter Weise zur Transformation von Pflanzen, insbesondere von Kulturpflanzen, wie z.B. von Tabakpflanzen, verwendet werden, indem
20 beispielsweise verwundete Blätter oder Blattstücke in einer Agrobakterienlösung gebadet und anschließend in geeigneten Medien kultiviert werden. Die Transformation von Pflanzen durch Agrobakterien ist unter anderem bekannt aus F.F. White, Vectors for Gene Transfer in Higher Plants; in Transgenic Plants, Vol. 1, Engineering and Utilization, herausgegeben von S.D. Kung und R. Wu,
25 Academic Press, 1993, S. 15-38 und aus S.B. Gelvin, Molecular Genetics of T-DNA Transfer from Agrobacterium to Plants, gleichfalls in Transgenic Plants, S. 49-78. Aus den transformierten Zellen der verwundeten Blätter bzw. Blattstücke können in bekannter Weise transgene Pflanzen regeneriert werden, die ein in die
30 erfindungsgemäße Expressionskassette integriertes Gen für die Expression eines Polypeptides mit Herbizid-bindenden Eigenschaften enthalten.

35 Zur Transformation einer Wirtspflanze mit einer für ein Herbizid-bindendes Polypeptid codierenden DNA wird eine erfindungsgemäße Expressionskassette als Insertion in einen rekombinanten Vektor eingebaut, dessen Vektor-DNA zusätzliche funktionelle Regulationssignale, beispielsweise Sequenzen für Replikation oder Integration enthält. Geeignete Vektoren sind unter anderem in
40 "Methods in Plant Molecular Biology and Biotechnology" (CRC Press), Kap. 6/7, S.71-119 (1993) beschrieben.

Unter Verwendung der oben zitierten Rekombinations- und
45 Klonierungstechniken können die erfindungsgemäßen Expressionskassetten in geeignete Vektoren kloniert werden, die ihre Vermehrung, beispielsweise in E. coli, ermöglichen. Geeignete Klonierung,

11

rungsvektoren sind u.a. pBR332, pUC-Serien, M13mp-Serien und pACYC184. Besonders geeignet sind binäre Vektoren, die sowohl in *E. coli* als auch in Agrobakterien replizieren können, wie z.B. pBin19 (Bevan et al. (1980) Nucl. Acids Res. 12, 8711).

5

Ein weiterer Gegenstand der Erfindung betrifft die Verwendung einer erfindungsgemäßen Expressionskassette zur Transformation von Pflanzen, Pflanzenzellen, -geweben oder Pflanzenteilen. Vorzugsweise ist Ziel der Verwendung die Vermittlung von

10 Resistenz gegen Inhibitoren pflanzlicher Enzyme.

Dabei kann je nach Wahl des Promotors die Expression spezifisch in den Blättern, in den Samen oder anderen Teilen der Pflanze erfolgen. Solche transgenen Pflanzen, deren Vermehrungsgut sowie
 15 deren Pflanzenzellen, -gewebe oder -teile sind ein weiterer Gegenstand der vorliegenden Erfindung.

Die Übertragung von Fremdgenen in das Genom einer Pflanze wird als Transformation bezeichnet. Es werden dabei die beschriebenen
 20 Methoden zur Transformation und Regeneration von Pflanzen aus Pflanzengeweben oder Pflanzenzellen zur transienten oder stabilen Transformation genutzt. Geeignete Methoden sind die Protoplasten-transformation durch Polyethylenglykol-induzierte DNA-Aufnahme, der biolistische Ansatz mit der Genkanone, die Elektroporation,
 25 die Inkubation trockener Embryonen in DNA-haltiger Lösung, die Mikroinjektion und der Agrobacterium-vermittelte Gentransfer. Die genannten Verfahren sind beispielsweise in B. Jenes et al., Techniques for Gene Transfer, in: Transgenic Plants, Vol. 1, Engineering and Utilization, herausgegeben von S.D. Kung und R. Wu,
 30 Academic Press (1993) 128-143 sowie in Potrykus Annu.Rev.Plant Physiol.Plant Molec.Biol. 42 (1991) 205-225 beschrieben. Vorzugsweise wird das zu exprimierende Konstrukt in einen Vektor kloniert, der geeignet ist, *Agrobacterium tumefaciens* zu transformieren, beispielsweise pBin19 (Bevan et al., Nucl. Acids Res.
 35 12 (1984) 8711).

Mit einer erfindungsgemäßen Expressionskassette transformierte Agrobakterien können dann in bekannter Weise zur Transformation von Pflanzen, insbesondere von Kulturpflanzen, wie Getreide,
 40 Mais, Soja, Reis, Baumwolle, Zuckerrübe, Canola, Sonnenblume, Flachs, Kartoffel, Tabak, Tomate, Raps, Alfalfa, Salat und den verschiedenen Baum-, Nuß- und Weinspezies, verwendet werden, z.B. indem verwundete Blätter oder Blattstücke in einer Agrobakterienlösung gebadet und anschließend in geeigneten Medien kultiviert
 45 werden.

12

Funktionell äquivalente Sequenzen, die für ein Herbizid-bindendes Polypeptid codieren, sind erfindungsgemäß solche Sequenzen, welche trotz abweichender Nucleotidsequenz noch die gewünschten Funktionen besitzen. Funktionelle Äquivalente umfassen somit natürlich vorkommende Varianten der hierin beschriebenen Sequenzen sowie künstliche, z.B. durch chemische Synthese erhaltene, an die Codon-Usage einer Pflanze angepasste, künstliche Nucleotid-Sequenzen.

- 10 Unter einem funktionellen Äquivalent versteht man insbesondere auch natürliche oder künstliche Mutationen einer ursprünglich isolierten das Herbizid-bindende Polypeptid codierenden Sequenz, welche weiterhin die gewünschte Funktion zeigen. Mutationen umfassen Substitutionen, Additionen, Deletionen, Vertauschungen
- 15 oder Insertionen eines oder mehrerer Nukleotidreste. Somit werden beispielsweise auch solche Nucleotidsequenzen durch die vorliegende Erfindung mit umfaßt, welche man durch Modifikation dieser Nucleotidsequenz erhält. Ziel einer solchen Modifikation kann z.B. die weitere Eingrenzung der darin enthaltenen codierenden
- 20 Sequenz oder z.B. auch die Einfügung weiterer Restriktionsenzym-Schnittstellen sein.

Funktionelle Äquivalente sind auch solche Varianten, deren Funktion, verglichen mit dem Ausgangsgen bzw. Genfragment, abge-

25 schwächt oder verstärkt ist.

Außerdem sind artifizielle DNA-Sequenzen geeignet, solange sie, wie oben beschrieben, die gewünschte Resistenz gegenüber Herbiziden induzieren. Solche artifiziellen DNA-Sequenzen können

30 beispielsweise durch Rückübersetzung mittels Molecular Modelling konstruierter Proteine, die Herbizid-bindende Aktivität aufweisen oder durch in vitro-Selektion ermittelt werden. Besonders geeignet sind kodierende DNA-Sequenzen, die durch Rückübersetzung einer Polypeptidsequenz gemäß der für die Wirtspflanze spezifischen Codon-Nutzung erhalten wurden. Die spezifische Codon-Nut-

35 zung kann ein mit pflanzen genetischen Methoden vertrauter Fachmann durch Computerauswertungen anderer, bekannter Gene der zu transformierenden Pflanze leicht ermitteln.

40 Als weitere erfindungsgemäße geeignete äquivalente Nukleinsäure-Sequenzen sind zu nennen Sequenzen, welche für Fusionsproteine kodieren, wobei Bestandteil des Fusionsproteins ein nicht-pflanzliches Herbizid-bindendes Polypeptid oder ein funktionell äquivalenter Teil davon ist. Der zweite Teil des Fusionsproteins kann

45 z.B. ein weiteres Polypeptid mit enzymatischer Aktivität sein oder eine antigene Polypeptidsequenz mit deren Hilfe ein Nachweis auf scFvs Expression möglich ist (z.B. myc-tag oder his-tag).

13

Bevorzugt handelt es sich dabei jedoch um eine regulative Proteinsequenz, wie z.B. ein Signal- oder Transitpeptid, das das Polypeptid mit Herbizid-bindenden Eigenschaften an den gewünschten Wirkort leitet.

5

Gegenstand der Erfindung sind aber auch die erfindungsgemäß erzeugten Expressionsprodukte, sowie Fusionsproteine aus einem Transitpeptid und einem Polypeptid mit Herbizid-bindenden Eigenschaften.

10

Resistenz bzw. Toleranz bedeutet im Rahmen der vorliegenden Erfindung die künstlich erworbene Widerstandsfähigkeit gegen die Wirkung pflanzlicher Enzym-Inhibitoren. Sie umfaßt die partielle und, insbesondere, die vollständige Unempfindlichkeit gegenüber

15 diesen Inhibitoren für die Dauer mindestens einer Pflanzengeneration.

Der primäre Wirkort von Herbiziden ist im allgemeinen das Blattgewebe, so daß eine blattspezifische Expression des exogenen Her-

20 bizid-bindenden Polypeptides ausreichenden Schutz bieten kann. Es ist jedoch naheliegend, daß die Wirkung eines Herbizids nicht auf das Blattgewebe beschränkt sein muß, sondern auch in allen übrigen Teilen der Pflanze gewebespezifisch erfolgen kann.

25 Darüberhinaus ist eine konstitutive Expression des exogenen Herbizid-bindenden Polypeptides von Vorteil. Andererseits kann aber auch eine induzierbare Expression wünschenswert erscheinen.

Die Wirksamkeit des transgen exprimierten Polypeptides mit Herbizid-bindenden Eigenschaften kann beispielsweise in vitro durch
30 Sproßmeristemvermehrung auf Herbizid-haltigem Medium über abgestufte Konzentrationsreihen oder über Samenkeimungstests ermittelt werden. Zudem kann eine in Art und Höhe veränderte Herbizidverträglichkeit einer Testpflanze in Gewächshausversuchen getestet werden.
35

Gegenstand der Erfindung sind außerdem transgene Pflanzen, transformiert mit einer erfindungsgemäßen Expressionskassette, sowie transgene Zellen, Gewebe, Teile und Vermehrungsgut solcher Pflanzen.
40 zen. Besonders bevorzugt sind dabei transgene Kulturpflanzen, wie z.B. Getreide, Mais, Soja, Reis, Baumwolle, Zuckerrübe, Canola, Sonnenblume, Flachs, Kartoffel, Tabak, Tomate, Raps, Alfalfa, Salat und die verschiedenen Baum-, Nuß- und Weinspecies.

45 Die transgenen Pflanzen, Pflanzenzellen, -gewebe oder -teile können mit einem Wirkstoff, der die pflanzlichen Enzyme inhibiert, behandelt werden, wodurch die nicht erfolgreich transformierten

14

Pflanzen, -zellen, -gewebe oder Pflanzenteile absterben. Beispiele für geeignete Wirkstoffe sind insbesondere 5-(2-Chlor-4-(trifluormethyl)phenoxy)-2-nitrobenzoesäure (Acifluorfen) und 7-Chlor-3-methylchinolin-8-carbonsäure (Quinmerac), sowie Metabolite und funktionelle Derivate dieser Verbindungen. Die in die erfindungsgemäßen Expressionskassetten insertierte, für ein Polypeptid mit Herbizid-bindenden Eigenschaften codierende DNA, kann somit auch als Selektionsmarker verwendet werden.

- 10 Insbesondere bei Kulturpflanzen bietet die vorliegende Erfindung den Vorteil daß nach Induktion einer selektiven Resistenz der Kulturpflanze gegenüber pflanzlichen Enzym-Inhibitoren diese Inhibitoren als spezifische Herbizide gegen nicht resistente
- 15 Pflanzen eingesetzt werden können. Als nicht-limitierende Beispiele für derartige Inhibitoren können genannt werden die folgenden herbiziden Verbindungen aus den Gruppen b1 - b41 :

b1 1,3,4-Thiadiazolen:
20 buthidazole, cyprazole

b2 Amide:
allidochlor (CDAA), benzoylprop-ethyl, bromobutide, chlort-
hiamid, dimepiperate, dimethenamid, diphenamid, etobenzanid
25 (benzchlomet), flamprop-methyl, fosamin, isoxaben, monalide, naptalame, pronamid (propyzamid), propanil

b3 Aminophosphorsäuren:
bilanafos, (bialaphos), buminafos, glufosinate-ammonium, gly-
30 phosate, sulfosate

b4 Aminotriazolen:
amitrol

35 b5 Anilide:
anilofos, mefenacet

b6 Aryloxyalkansäuren:
2,4-D, 2,4-DB, clomeprop, dichlorprop, dichlorprop-P, dich-
40 lorprop-P (2,4-DP-P), fenoprop (2,4,5-TP), fluoroxyppyr, MCPA, MCPB, mecoprop, mecoprop-P, napropamide, napropanilide, tri-
clopyr

b7 Benzoessäuren:
45 chloramben, dicamba

15

- b8 Benzothiadiazinonen:
bentazon
- b9 Bleacher:
5 clomazone (dimethazone), diflufenican, fluorochloridone, flupoxam, fluridone, pyrazolate, sulcotrione (chlormesulone)
- b10 Carbamaten:
10 asulam, barban, butylate, carbetamid, chlorbufam, chlorpropham, cycloate, desmedipham, diallate, EPTC, esprocarb, molinate, orbencarb, pebulate, phenisopham, phenmedipham, propham, prosulfocarb, pyributicarb, sulfallate (CDEC), terbutcarb, thiobencarb (benthicarb), tiocarbazil, triallate, ver-nolate
- 15 b11 Chinolinsäuren:
quinclorac, quinmerac
- b12 Chloracetaniliden:
20 acetochlor, alachlor, butachlor, butenachlor, diethatyl ethyl, dimethachlor, metazachlor, metolachlor, pretilachlor, propachlor, prynachlor, terbuchlor, thenylchlor, xylachlor
- b13 Cyclohexenonen:
25 alloxydim, caloxydim, clethodim, cloproxydim, cycloxydim, sethoxydim, tralkoxydim, 2-{1-[2-(4-Chlorphenoxy)propyloxyimino]butyl}-3-hydroxy-5-(2H-tetrahydrothiopyran-3-yl)-2-cyclohexen-1-on
- 30 b14 Dichlorpropionsäuren:
dalapon
- b15 Dihydrobenzofurane:
ethofumesate
- 35 b16 Dihydrofuran-3-one:
flurtamone
- b17 Dinitroaniline:
40 benefin, butralin, dinitramin, ethalfluralin, fluchloralin, isopropalin, nitralin, oryzalin, pendimethalin, prodiamine, profluralin, trifluralin
- b18 Dinitrophenole:
45 bromofenoxim, dinoseb, dinoseb-acetat, dinoterb, DNOC

- b19 Diphenylether:
acifluorfen-sodium, acionifen, bifenox, chlornitrofen (CNP),
difenoxuron, ethoxyfen, fluorodifen, fluoroglycofen-ethyl,
fomesafen, furyloxyfen, lactofen, nitrofen, nitrofluorfen,
oxyfluorfen
- b20 Dipyridylene:
cyperquat, difenzoquat-methylsulfat, diquat, paraquat di-
chlorid
- b21 Harnstoffe:
benzthiazuron, buturon, chlorbromuron, chloroxuron, chlorto-
luron, cumyluron, dibenzyluron, cycluron, dimefuron, diuron,
dymron, ethidimuron, fenuron, fluormeturon, isoproturon,
isouron, karbutilat, linuron, methabenzthiazuron, metobenzu-
ron, metoxuron, monolinuron, monuron, neburon, siduron, tebu-
thiuron, trimeturon
- b22 Imidazole:
isocarbamid
- b23 Imidazolinone:
imazamethapyr, imazapyr, imazaquin, imazethabenz-methyl (ima-
zame), imazethapyr
- b24 Oxadiazole:
methazole, oxadiargyl, oxadiazon
- b25 Oxirane:
tridiphane
- b26 Phenole:
bromoxynil, ioxynil
- b27 Phenoxyphenoxypropionsäureester:
clodinafop, cyhalofop-butyl, diclofop-methyl, fenoxaprop-
ethyl, fenoxaprop-p-ethyl, fenthiapropethyl, fluazifop-butyl,
fluazifop-p-butyl, haloxyfop-ethoxyethyl, haloxyfop-methyl,
haloxyfop-p-methyl, isoxapyrifop, propaquizafop, quizalofop-
ethyl, quizalofop-p-ethyl, quizalofop-tefuryl
- b28 Phenyllessigsäuren:
chlorfenac (fenac)
- b29 Phenylpropionsäuren:
chlorophenprop-methyl

17

- b30 Protoporphyrinogen-IX-Oxydase-Hemmer:
benzofenap, cinidon-ethyl, flumiclorac-pentyl, flumioxazin,
flumipropyn, flupropacil, fluthiacet-methyl, pyrazoxyfen,
sulfentrazone, thidiazimin
- 5
- b31 Pyrazole:
nipyraclufen
- b32 Pyridazine:
10 chloridazon, maleic hydrazide, norflurazon, pyridate
- b33 Pyridincarbonsäuren:
clopyralid, dithiopyr, picloram, thiazopyr
- 15 b34 Pyrimidylethern:
pyrithiobac-säure, pyrithiobac-sodium, KIH-2023, KIH-6127
- b35 Sulfonamide:
flumetsulam, metosulam
- 20
- b36 Sulfonylharnstoffe:
amidosulfuron, azimsulfuron, bensulfuron-methyl, chlorimuron-
ethyl, chlorsulfuron, cinosulfuron, cyclosulfamuron, ethamet-
sulfuron methyl, ethoxysulfuron, flazasulfuron, halosulfuron-
25 methyl, imazosulfuron, metsulfuron-methyl, nicosulfuron, pri-
misulfuron, prosulfuron, pyrazosulfuron-ethyl, rimsulfuron,
sulfometuron-methyl, thifensulfuron-methyl, triasulfuron,
tribenuron-methyl, triflusulfuron-methyl
- 30 b37 Triazine:
ametryn, atrazin, aziprotryn, cyanazine, cyprazine, desme-
tryn, dimethamethryn, dipropetryn, eglinazin-ethyl, hexazi-
non, procyazine, prometon, prometryn, propazin, secbumeton,
simazin, simetryn, terbumeton, terbutryn, terbutylazin, trie-
35 tazin
- b38 Triazinone:
ethiozin, metamitron, metribuzin
- 40 b39 Triazolcarboxamide:
triazofenamid
- b40 Uracile:
bromacil, lenacil, terbacil
- 45

b41 Verschiedene:

- benazolin, benfuresate, bensulide, benzofluor, butamifos, ca-
fenstrole, chlorthal-dimethyl (DCPA), cinmethylin, dichlobe-
nil, endothall, fluorbentranil, mefluidide, perfluidone, pi-
perophos

Funktionell äquivalente Derivate pflanzlicher Enzym-Inhibitoren
besitzen ein vergleichbares Wirkungsspektrum wie die konkret ge-
nannten Substanzen, bei niedrigerer, gleicher oder höherer inhi-
bitorischer Aktivität (z.B. ausgedrückt in g Inhibitor pro Hektar
Anbaufläche, erforderlich zur vollständigen Unterdrückung des
Wachstums nicht-resistenter Pflanzen).

Die Erfindung wird durch die nun folgenden Beispiele erläutert,
ist aber nicht auf diese beschränkt:

Allgemeine Klonierungsverfahren

Die im Rahmen der vorliegenden Erfindung durchgeführten Klonie-
rungsschritte wie z.B. Restriktionsspaltungen, Agarose-Gel-
elektrophorese, Reinigung von DNA-Fragmenten, Transfer von Nu-
kleinsäuren auf Nitrozellulose und Nylonmembranen, Verknüpfen von
DNA-Fragmenten, Transformation von E. coli Zellen, Anzucht von
Bakterien, Vermehrung von Phagen und Sequenzanalyse rekombinanter
DNA wurden wie bei Sambrook et al. (1989) Cold Spring Harbor La-
boratory Press; ISBN 0-87969-309-6) beschrieben durchgeführt.

Die im folgenden verwendeten Bakterienstämme (E. coli, XL-I Blue)
wurden von Stratagene bezogen. Der zur Pflanzentransformation
verwendete Agrobakterienstamm (Agrobacterium tumefaciens, C58C1
mit dem Plasmid pGV2260 oder pGV3850kan) wurde von Deblaere et
al. beschrieben (Nucl. Acids Res. 13 (1985) 4777). Alternativ
können auch der Agrobakterienstamm LBA4404 (Clontech) oder andere
geeignete Stämme eingesetzt werden. Zur Klonierung wurden die
Vektoren pUC19 (Yanish-Perron, Gene 33 (1985), 103-119) pBlues-
cript SK- (Stratagene), pGEM-T (Promega), pZero (Invitrogen),
pBin19 (Bevan et al., Nucl. Acids Res. 12 (1984) 8711-8720) und
pBinAR (Höfgen und Willmitzer, Plant Science 66 (1990) 221-230)
benutzt.

40

Sequenzanalyse rekombinanter DNA

Die Sequenzierung rekombinanter DNA-Moleküle erfolgte mit einem
Laserfluoreszenz-DNA-Sequenzierer der Firma Pharmacia nach der
Methode von Sanger (Sanger et al., Proc. Natl. Acad. Sci. USA
74 (1977), 5463-5467).

19

Erzeugung pflanzlicher Expressionskassetten

In das Plasmid pBin19 (Bevan et al., Nucl. Acids Res. 12, 8711 (1984)) wurde ein 35S CaMV Promotor als EcoRI-KpnI-Fragment
 5 (entsprechend den Nukleotiden 6909-7437 des Cauliflower-Mosaik-Virus (Franck et al. Cell 21 (1980) 285) inseriert. Das Polyadenylierungssignal des Gens 3 der T-DNA des Ti-Plasmides pTiACH5 (Gielen et al., EMBO J. 3 (1984) 835), Nukleotide 11749-11939 wurde als PvuII-HindIII-Fragment isoliert und nach Addition von
 10 SphI-Linkern an die PvuII-Schnittstelle zwischen die SphI-HindIII Schnittstelle des Vektors kloniert. Es entstand das Plasmid pBinAR (Höfgen und Willmitzer, Plant Science 66 (1990) 221-230).

Anwendungsbeispiele

15

Beispiel 1

Da Herbizide nicht immunogen sind, müssen sie an ein Trägermaterial wie z.B. KLH gekoppelt werden. Befindet sich eine reaktive Gruppe im Molekül, kann diese Kopplung direkt erfolgen, ansonsten wird während der Synthese des Herbizides eine funktionelle Gruppe eingeführt oder eine reaktive Vorstufe während der Synthese ausgesucht, um diese Moleküle in einem einfachen Reaktionsschritt an das Trägermolekül zu koppeln. Beispiele für Kopplungen sind bei Miroslavic Ferencik in "Handbook of Immunochemistry", 1993, Chapman & Hall, im Kapitel Antigene, Seite 20 -
 25 49 beschrieben.

Durch wiederholte Injektion dieses modifizierten Trägermoleküls (Antigens) werden z.B. Balb/c-Mäuse immunisiert. Sobald im Serum genügend Antikörper mit Bindung an das Antigen im ELISA (enzyme linked immuno sorbent assay) nachweisbar sind, werden die Milzzellen dieser Tiere entnommen und mit Myelomzellen fusioniert um Hybride zu kultivieren. Im ELISA wird zusätzlich als Antigen
 30 "Herbizid-modifiziertes BSA" verwendet, um die gegen das Hapten gerichtete Immunantwort von der KLH-Antwort zu unterscheiden.

Die Herstellung von monoklonalen Antikörpern erfolgt in Anlehnung an bekannte Methoden, wie z.B. beschrieben in "Practical Immunology", Leslie Hudson und Frank Hay, Blackwell Scientific
 40 Publications, 1989 oder in "Monoclonal Antibodies: Principles and Practice", James Goding, 1983, Academic Press, Inc., oder in "A practical guide to monoclonal antibodies", J.Liddell und A. Cryer, 1991, John Wiley & Sons; oder Achim Möller und Franz Emling
 45 "Monoklonale Antikörper gegen TNF und deren Verwendung". Europäische Patentschrift EP-A260610.

20

Beispiel 2

Ausgangspunkt der Untersuchung war ein monoklonaler Antikörper der spezifisch das Herbizid Quinmerac erkennt und der außerdem
5 eine hohe Bindungsaffinität aufweist. Die selektionierte Hybridomazelllinie ist dadurch charakterisiert, daß die sekretierten, gegen das Herbizid-Antigen Quinmerac gerichteten monoklonalen Antikörper eine hohe Affinität aufweisen und die spezifischen Sequenzen der Immunglobuline verfügbar sind (Berek, C. et al., Nature
10 316, 412-418 (1985)). Dieser monoklonale Antikörper gegen Quinmerac war Ausgangspunkt für die Konstruktion des Einketten-Antikörperfragmentes (scFv-antiQuinmerac).

Zunächst wurde mRNA aus den Hybridomzellen isoliert und in cDNA
15 umgeschrieben. Diese cDNA diente als Matrize für die Amplifikation der variablen Immunglobulingene VH und VK mit den spezifischen Primern VH1 BACK und VH FOR-2 für die schwere Kette sowie VK2 BACK und MJK5 FON X für die leichte Kette (Clackson et al., Nature 352, 624-628 (1991)). Die isolierten variablen Immunglobulin-
20 line waren Ausgangspunkt für die Konstruktion eines Einketten-Antikörperfragmentes (scFv-antiQuinmerac). Bei der nachfolgenden Fusions-PCR wurden drei Komponenten VH, VK und ein Linkerfragment in einem PCR-Reaktionsansatz vereinigt und das scFv-antiQuinmerac amplifiziert (Abb. 3).

25

Die funktionelle Charakterisierung (Antigenbindungsaktivität) des konstruierten scFv-antiQuinmerac-Gens erfolgte nach Expression in einem bakteriellen System. Das scFv-antiQuinmerac wurde dazu nach der Methode von Hoogenboom, H.R. et al., Nucleic Acids
30 Research, 19, 4133-4137 (1991) als lösliches Antikörperfragment in E.coli synthetisiert. Die Aktivität und die Spezifität des konstruierten Antikörperfragmentes wurden in ELISA-Tests überprüft (Abb.4).

35 Um eine samenspezifische Expression des Antikörperfragmentes in Tabak zu ermöglichen, wurde das scFv-antiQuinmerac Gen stromabwärts vom LeB4-Promotor kloniert. Der aus Vicia faba isolierte LeB4-Promotor zeigt eine streng samenspezifische Expression von verschiedenen Fremdgenen in Tabak (Bäumlein, H. et al., Mol. Gen.
40 Genet. 225, 121-128 (1991)). Durch Transport des scFv-antiQuinmerac Polypeptides in das endoplasmatische Retikulum wurde eine stabile Akkumulation hoher Antikörperfragmentmengen erreicht. Das scFv-antiQuinmerac Gen wurde zu diesem Zweck mit einer Signalpeptidsequenz, die den Eintritt in das endoplasmatische Retikulum
45 und dem ER-Retentionssignal SEKDEL, das ein Verbleiben im ER gewährleistet (Wandelt et al., 1992), fusioniert (Abb. 5).

21

- Die konstruierte Expressionskassette wurde in den binären Vektor pGSGLuc 1 (Saito et al., 1990) kloniert und durch Elektroporation in den *Agrobacterium*-Stamm EHA 101 transferiert. Rekombinante *Agrobacterien*klone wurden für die nachfolgende Transformation von *Nicotiana tabacum* verwendet. Pro Konstrukt wurden 70-140 Tabakpflanzen regeneriert. Von den regenerierten transgenen Tabakpflanzen wurden nach Selbstbefruchtung Samen verschiedener Entwicklungsstadien geerntet. Von diesen Samen wurden die löslichen Proteine nach Extraktion in einem wässrigen Puffersystem erhalten. Die Analyse der transgenen Pflanzen zeigt, daß durch die Fusion des scFv-antiQuinmerac Gens mit der DNA-Sequenz des ER-Retentionssignals SEKDEL eine maximale Akkumulation von 1,9 % scFv-antiQuinmerac Protein im reifen Samen erzielt werden konnte.
- 15 Das konstruierte scFv-antiQuinmerac Gen hatte eine Größe von ca. 735 bp. Die variablen Domänen wurden in der Reihenfolge VH-L-VL miteinander fusioniert.

- Die spezifische Selektivität wurde in den Extrakten der reifen Tabaksamen mit einem direkten ELISA bestimmt. Die dabei erhaltenen Werte zeigen deutlich, daß die Proteinextrakte funktionell aktive Antikörperfragmente enthalten.

Beispiel 3

- 25 Samenspezifische Expression und Anreicherung von Einketten-Antikörperfragmenten im endoplasmatischen Retikulum von Zellen transgener Tabaksamen kontrolliert durch den USP- Promotor.
- 30 Ausgangspunkt der Untersuchungen war ein Einzelketten-Antikörperfragment gegen das Herbizid Quinmerac (scFv-anti Quinmerac). Die funktionelle Charakterisierung (Antigenbindungsaktivität) dieses konstruierten scFv-anti- Quinmerac Genes erfolgte nach Expression in einem bakteriellen System und nach Expression in Tabakblättern. Die Aktivität und die Spezifität des konstruierten Antikörperfragmentes wurde in ELISA-Tests überprüft.

- Um eine samenspezifische Expression des Antikörperfragmentes in Tabak zu ermöglichen, wurde das scFv-antiQuinmerac Gen stromaufwärts vom USP-Promotor kloniert. Der aus *Vicia faba* isolierte USP-Promotor zeigt eine streng samenspezifische Expression von verschiedenen Fremdgenen in Tabak (Fiedler, U. et al., Plant Mol. Biol. 22, 669-679 (1993)). Durch Transport des scFv-anti-Quinmerac Polypeptides in das endoplasmatische Retikulum wurde eine stabile Akkumulation hoher Antikörperfragmentmengen erreicht. Das scFv-antiQuinmerac Gen wurden zu diesem Zweck mit einer Signalpeptidsequenz, die den Eintritt in das endoplasmatische

22

sche Retikulum und dem ER-Retentionssignal SEKDEL, das ein Verbleiben im ER gewährleistet (Wandelt et al., 1992), fusioniert (Abb. 1).

- 5 Die konstruierte Expressionskassette wurde in den binären Vektor pGSGLUC1 (Saito et al., 1990) kloniert und durch Elektroporation in den Agrobakterium-Stamm EHA 101 transferiert. Rekombinante Agrobaktienklone wurden für die nachfolgende Transformation von *Nicotiana tabacum* verwendet. Von den regenerierten transgenen Tabakpflanzen wurden nach Selbstbefruchtung Samen verschiedener Entwicklungsstadien geerntet. Von diesen Samen wurden die löslichen Proteine nach Extraktion in einem wässrigen Puffersystem erhalten. Die Analyse der transgenen Pflanzen zeigt, daß durch die Fusion des scFv-antiAcifluorfen Gens mit der DNA-Sequenz des ER-Retentionssignals SEKDEL unter Kontrolle des USP-Promotors bereits ab Tag 10 der Samenentwicklung Einketten-Antikörperfragmente mit Bindeaffinität für Quinmerac synthetisiert wurden.

Beispiel 4

20

- Um eine ubiquitäre Expression des Antikörperfragmentes in der Pflanze, speziell in Blättern, zu erreichen, wurde das scFv-anti-Quinmerac -Gen stromabwärts vom CaMV 35 S-Promotor kloniert. Dieser starke konstitutive Promotor vermittelt eine Expression von Fremdgenen in nahezu allen pflanzlichen Geweben (Benfey und Chua, Science 250 (1990), 956 - 966). Durch Transport des scFv-anti-Quinmerac Proteins in das endoplasmatische Retikulum wurde eine stabile Akkumulation hoher Antikörperfragmentmengen im Blattmaterial erreicht. Das scFv-antiQuinmerac Gen wurde zunächst mit einer Signalpeptidsequenz, die den Eintritt in das endoplasmatische Retikulum und dem ER-Retentionssignal KDEL, das ein Verbleiben im ER gewährleistet (Wandelt et al., Plant J. 2(1992), 181 - 192) fusioniert. Die konstruierte Expressionskassette wurde in den binären Vektor pGSGLUC 1 (Saito et al., Plant Cell Rep. 8(1990), 718 - 721) kloniert und durch Elektroporation in den Agrobakterium -Stamm EHA 101 transferiert. Rekombinante Agrobaktienklone wurden für die nachfolgende Transformation von *Nicotiana tabacum* verwendet. Es wurden ungefähr 100 Tabakpflanzen regeneriert. Von den regenerierten transgenen Tabakpflanzen wurde Blattmaterial verschiedenener Entwicklungsstufen entnommen. Von diesem Blattmaterial wurden die löslichen Proteine nach Extraktion in einem wässrigen Puffersystem erhalten. Nachfolgende Analysen (Western-Blot-Analysen und ELISA-Tests) zeigten, daß in den Blättern eine maximale Akkumulation von größer 2 % an biologisch aktivem, antigenbindendem scFv-antiQuinmerac Polypeptid erzielt werden konnte. Die hohen Expressionswerte wurden in ausgewachse-

nen grünen Blättern ermittelt, aber auch in seneszenten Blattmaterial konnte das Antikörperfragment nachgewiesen werden.

Beispiel 5

- 5 PCR-Amplifikation eines Fragmentes der cDNA codierend für den Ein-Ketten-Antikörper gegen Acifluorfen bzw. Quinmerac mithilfe synthetischer Oligonukleotide.
- 10 Die PCR-Amplifikation der Ein-Ketten-Antikörper cDNA wurde in einem DNA-Thermal Cycler der Firma Perkin Elmer durchgeführt. Die Reaktionsgemische enthielten 8 ng/ μ l einzelsträngige Matrizen-cDNA, 0,5 μ M der entsprechenden Oligonukleotide, 200 μ M Nukleotide (Pharmacia), 50 mM KCl, 10 mM Tris-HCl (pH 8,3 bei 25°C, 1,5mM MgCl₂) und 0.02 U/ μ l Taq Polymerase (Perkin Elmer). Die Amplifikationsbedingungen wurden wie folgt eingestellt:

Anlagerungstemperatur:	45°C
Denaturierungstemperatur:	94°C,
20 Elongationstemperatur:	72°C,
Anzahl der Zyklen:	40

- Es resultierte ein Fragment von ca. 735 Basenpaaren, das in den Vektor pBluescript ligiert wurde. Mit dem Ligationsansatz wurde
- 25 E. coli XL-I Blue transformiert und das Plasmid amplifiziert. Zur Anwendung und Optimierung der Polymerase Kettenreaktion siehe: Innis et al., 1990, PCR Protocols, a Guide to Methods and Applications, Academic Press.

30 Beispiel 6

Herstellung transgener Tabakpflanzen, die eine cDNA codierend für einen Ein-Ketten-Antikörper mit Herbizid-bindenden Eigenschaften exprimieren.

- 35 Das Plasmid pGSGLUC 1 wurde in *Agrobacterium tumefaciens* C58C1:pGV2260 transformiert. Zur Transformation von Tabakpflanzen (*Nicotiana tabacum* cv. Samsun NN) wurde eine 1:50 Verdünnung einer Übernachtskultur einer positiv transformierten Agrobakterienkolonie in Murashige-Skoog Medium (Physiol. Plant. 15 (1962) 473 ff.) mit 2% Saccharose (2MS-Medium) benutzt. Blattscheiben steriler Pflanzen (zu je ca. 1 cm²) wurden in einer Petrischale mit einer 1:50 Agrobakterienverdünnung für 5-10 Minuten inkubiert. Es folgte eine 2-tägige Inkubation in Dunkelheit bei
- 45 25°C auf 2MS-Medium mit 0,8% Bacto-Agar. Die Kultivierung wurde nach 2 Tagen mit 16 Stunden Licht/8 Stunden Dunkelheit weitergeführt und in wöchentlichem Rhythmus auf MS-Medium mit 500 mg/l

24

Claforan (Cefotaxime-Natrium), 50 mg/l Kanamycin, 1 mg/l Benzylaminopurin (BAP), 0,2 mg/l Naphtylelessigsäure und 1,6 g/l Glukose weitergeführt. Wachsende Sprosse wurden auf MS-Medium mit 2% Saccharose, 250 mg/l Claforan und 0,8% Bacto-Agar überführt.

5

Beispiel 7

Stabile Akkumulation des Einketten-Antikörperfragmentes gegen das Herbizid Quinmerac im endoplasmatischen Reticulum.

10

Ausgangspunkt der Untersuchungen war ein in Tabakpflanzen exprimiertes Einketten-Antikörperfragment gegen das Herbizid Quinmerac(scFv-anti Quinmerac). Menge und Aktivität des synthetisierten scFv-antiQuinmerac Polypeptides wurden in Western-Blot-Analysen

15 und ELISA-Tests bestimmt.

Um eine Expression des scFv-antiQuinmerac-Gens im endoplasmatischen Retikulum zu ermöglichen, wurde das Fremdgen unter der Kontrolle des CaMV 53S-Promotors als eine Translationsfusion mit
20 dem LeB4-Signalpeptid (N-terminal) und dem ER-Retentionssignal KDEL (C-terminal) exprimiert. Durch Transport des scFv-antiQuinmerac Polypeptids in das endoplasmatische Retikulum wurde eine stabile Akkumulation hoher Mengen an aktivem Antikörperfragment erreicht. Nach Ernte des Blattmaterials wurden Stücke bei -20°C
25 eingefroren (1), lyophilisiert (2) oder bei Raumtemperatur getrocknet (3). Die löslichen Proteine wurden aus dem jeweiligen Blattmaterial durch Extraktion in einem wässrigen Puffer erhalten und das scFv-antiQuinmerac Polypeptid affinitätschromatographisch gereinigt. Gleiche Mengen an gereinigtem scFv-antiQuin-
30 merac Polypeptids (eingefroren, lyophilisiert und getrocknet) wurden für die Bestimmung der Aktivität des Antikörperfragmentes eingesetzt (Abb. 6). In Abb. 6 A ist die Antigenbindungsaktivität des aus frischen (1), lyophilisierten (2) und getrockneten
Blättern (3) gereinigten scFv-antiQuinmerac Polypeptides dargestellt. In Abb. 6 B sind die jeweiligen Mengen an scFv-antiQuinmerac Protein (etwa 100 ng), die für die ELISA-Analysen eingesetzt wurden, mittels Western-Blot-Analysen bestimmt. Die Größen der Proteinmolekulargewichtsstandards sind links dargestellt. Dabei wurden etwa gleiche Antigenbindungsaktivitäten festgestellt.

40

Beispiel 8

Zum Nachweis der Herbizid-Toleranz der ein Polypeptid mit Herbizid-bindenden Eigenschaften produzierenden transgenen Tabakpflanzen wurden diese mit unterschiedlichen Mengen Acifluorfen bzw.
45 Quinmerac behandelt. In allen Fällen konnte im Gewächshaus gezeigt werden, daß die ein scFv-antiAcifluorfen bzw. ein scFv-an-

25

tiQuinmerac exprimierenden Pflanzen im Vergleich zur Kontrolle
eine Toleranz gegenüber den entsprechenden Herbiziden zeigen.

5

10

15

20

25

30

35

40

45

Patentansprüche

1. Verfahren zur Herstellung von Herbizid-toleranten Pflanzen
5 durch Expression eines exogenen Herbizid-bindenden Polypeptids in den Pflanzen.
2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß es
10 sich bei dem exogenen Herbizid-bindenden Polypeptid um ein Einketten-Antikörperfragment handelt.
3. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß es
15 sich bei dem exogenen Herbizid-bindenden Polypeptid um einen kompletten Antikörper oder um ein davon abgeleitetes Fragment handelt.
4. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß es
20 sich bei dem Herbizid um 5-(2-Chlor-4-(trifluormethyl)phenoxy)-2-nitrobenzoesäure handelt.
5. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß es
sich bei dem Herbizid um 7-Chlor-3-methylchinolin-8-carbonsäure handelt.
- 25 6. Verfahren nach einem der Ansprüche 1 - 3, dadurch gekennzeichnet, daß es sich um mono- oder dikotyle Pflanzen handelt.
7. Verfahren nach Anspruch 6, dadurch gekennzeichnet, daß es
30 sich bei der Pflanze um Tabak handelt.
8. Verfahren nach einem der Ansprüche 1 - 7, dadurch gekennzeichnet, daß die Expression des exogenen Polypeptids konstitutiv in der Pflanze erfolgt.
- 35 9. Verfahren nach einem der Ansprüche 1 - 7, dadurch gekennzeichnet, daß die Expression des exogenen Polypeptids in der Pflanze induziert wird.
- 40 10. Verfahren nach einem der Ansprüche 1 - 7, dadurch gekennzeichnet, daß die Expression des exogenen Polypeptids in den Blättern der Pflanze erfolgt.
- 45 11. Verfahren nach einem der Ansprüche 1 - 7, dadurch gekennzeichnet, daß die Expression des exogenen Polypeptids in den Samen der Pflanze erfolgt.

Zeichn.

27

12. Expressionskassette für Pflanzen bestehend aus einem Promotor, einem Signalpeptid, einem Gen codierend für die Expression eines exogenen Herbizid-bindenden Polypeptids, einem ER-Retentionssignal und einem Terminator.

5

13. Expressionskassette nach Anspruch 12, dadurch gekennzeichnet, daß als konstitutiver Promotor der CaMV 35S-Promotor verwendet wird.
- 10 14. Expressionskassette nach Anspruch 12, dadurch gekennzeichnet, daß als zu exprimierendes Gen das Gen eines Einketten-Antikörperfragmentes eingesetzt wird.
- 15 15. Expressionskassette nach Anspruch 12, dadurch gekennzeichnet, daß als zu exprimierendes Gen das Gen oder Genfragment eines Herbizid-bindenden Polypeptides als Translationsfusion mit anderen funktionellen Proteinen wie zum Beispiel Enzymen, Toxinen, Chromophoren und Bindeproteinen eingesetzt wird.
- 20 16. Expressionskassette nach Anspruch 12, dadurch gekennzeichnet, daß das zu exprimierende Polypeptidgen aus einer Hybridomazelle oder mit Hilfe anderer rekombinanter Methoden - wie z.B. der Antikörper-Phage-Display Methode - gewonnen wird.
- 25 17. Verwendung der Expressionskassette nach Anspruch 12 zur Transformation von dicotylen oder monokotylen Pflanzen, die konstitutiv samen- oder blatt-spezifisch ein exogenes Herbizid-bindendes Polypeptid exprimieren.
- 30 18. Verwendung nach Anspruch 17, dadurch gekennzeichnet, daß man die Expressionskassette in einen Bakterienstamm transferiert und die entstandenen rekombinanten Klone zur Transformation von dicotylen oder monokotylen Pflanzen, die konstitutiv samen- oder blattspezifisch ein exogenes Herbizid-bindendes
- 35 Polypeptid exprimieren, verwendet.
19. Verwendung der Expressionskassette nach Anspruch 12 als Selektionsmarker.
- 40 20. Verwendung einer transformierten Pflanze wie nach Anspruch 18 oder 19 erhalten zur Herstellung eines Herbizid-bindenden Polypeptids.

45

28

21. Verfahren zur Transformation einer Pflanze durch Einbringen einer Gensequenz, die für ein Herbizid-bindendes Polypeptid codiert, in eine Pflanzenzelle, in Kallusgewebe, eine ganze Pflanze und Protoplasten von Pflanzenzellen.
- 5
22. Verfahren nach Anspruch 21, dadurch gekennzeichnet, daß die Transformation mit Hilfe eines Agrobacteriums insbesondere der Art *Agrobacterium tumefaciens* erfolgt.
- 10
23. Verfahren nach Anspruch 21, dadurch gekennzeichnet, daß die Transformation mit Hilfe der Elektroporation erfolgt.
24. Verfahren nach Anspruch 21, dadurch gekennzeichnet, daß die Transformation mit Hilfe der particle bombardment Methode erfolgt.
- 15
25. Herstellung eines Herbizid-bindenden Polypeptides durch Expression eines Gens codierend für ein derartiges Polypeptid in einer Pflanze bzw. Zellen einer Pflanze und anschließende Isolierung des Polypeptides.
- 20
26. Pflanze enthaltend eine Expressionskassette gemäß Anspruch 12, dadurch gekennzeichnet, daß die Expressionskassette Toleranz gegenüber einem Herbizid vermittelt.
- 25
27. Pflanze nach Anspruch 26, dadurch gekennzeichnet, daß sie tolerant gegenüber 5-(2-Chlor-4-(trifluormethyl)phenoxy)-2-nitrobenzoesäure ist.
- 30
28. Pflanze nach Anspruch 26, dadurch gekennzeichnet, daß sie tolerant gegenüber 7-Chlor-3-methylchinolin-8-carbonsäure ist.
29. Verfahren zur Bekämpfung von unerwünschtem Pflanzenwuchs in transgenen Herbizid-resistenten Kulturpflanzen dadurch gekennzeichnet, daß Herbizide eingesetzt werden, gegen die die Kulturpflanze Herbizid-bindende Polypeptide oder Antikörper bildet.
- 35
30. Herbizid-bindende Polypeptide bzw. Antikörper mit hoher Bindeaffinität zu 5-(2-Chlor-4-(trifluormethyl)phenoxy)-2-nitrobenzoesäure, dadurch gekennzeichnet, daß sie gemäß Anspruch 25 hergestellt werden.
- 40

29

31. Herbizid-bindende Polypeptide bzw. Antikörper mit hoher Bindeaffinität zu 7-Chlor-3-methylchinolin-8-carbonsäure, dadurch gekennzeichnet, daß sie gemäß Anspruch 25 hergestellt werden.

5

10

15

20

25

30

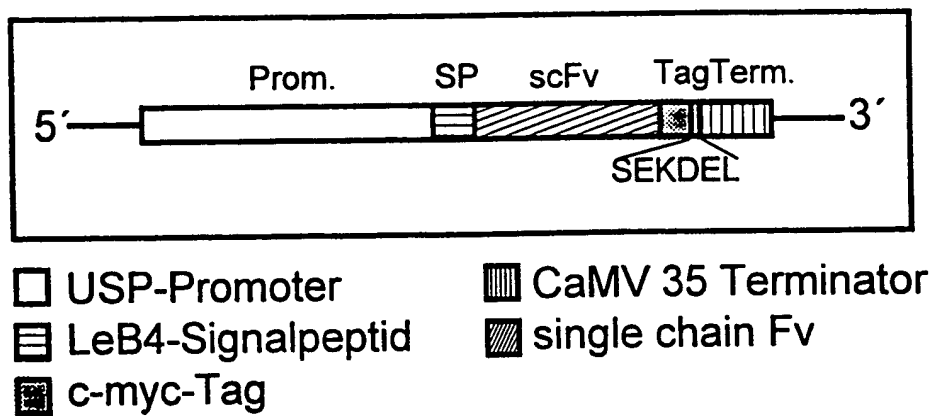
35

40

45

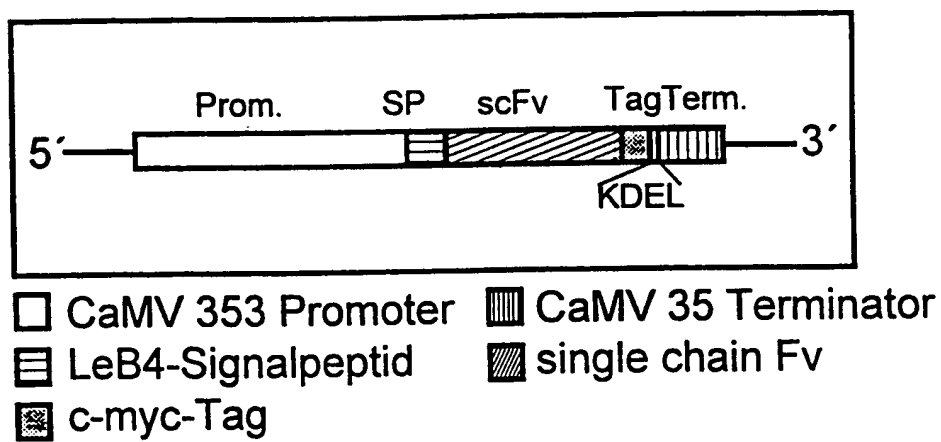
1/6

Fig. 1



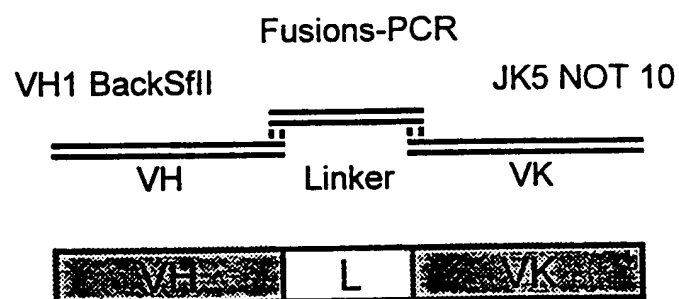
2/6

Fig. 2



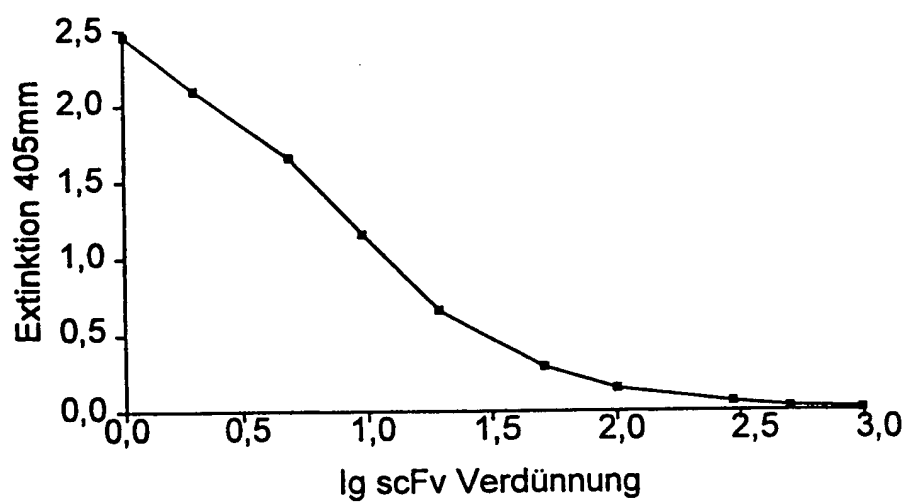
3/6

Fig. 3



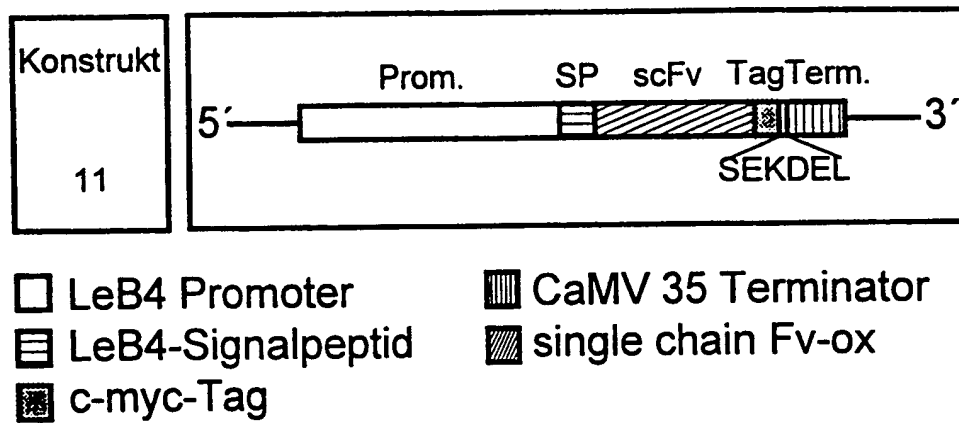
4/6

Fig. 4



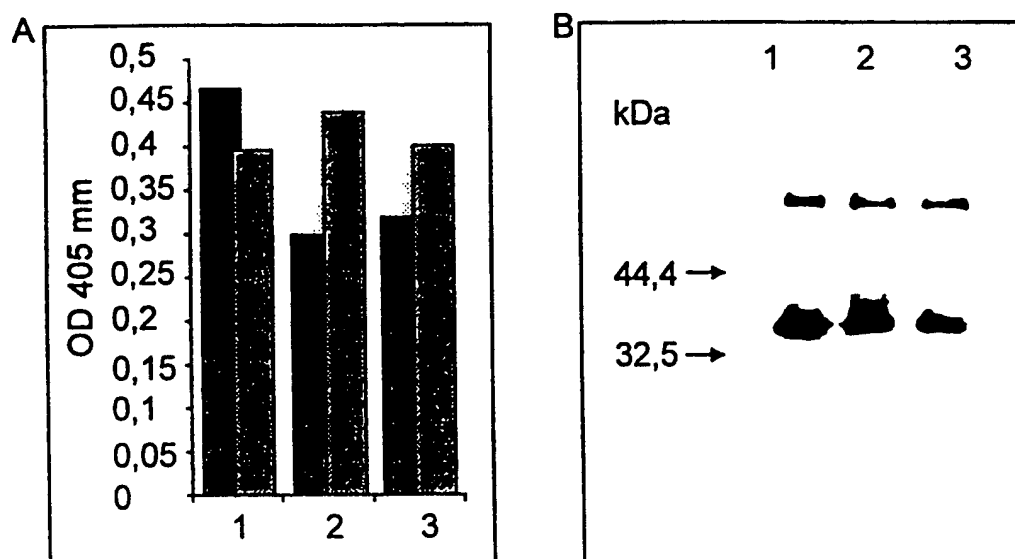
5/6

Fig. 5



6/6

Fig. 6



INTERNATIONAL SEARCH REPORT

In. ational Application No
PCT/EP 98/01731

A. CLASSIFICATION OF SUBJECT MATTER

IPC 6 C12N15/82 C07K16/44 C12N15/13 A01H5/00 A01N63/00

According to International Patent Classification(IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 C12N C07K A01H A01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ASBOTH, B.: "catalytically active antibodies for herbicide-resistant crop plant production ;antibody design for glyphosate inactivation; catalytic antibody; monoclonal antibody; abzyme may be used in the construction of a plant with glyphosate herbicide resistance (conference paper)" PLANT BIOTECHNOLGY, EC-HUNGARY WORKSHOP, 1991, page 21 XP002074579 see the whole document	1,3,6,8, 10,13, 21-24,29
X	SWAIN W F: "ANTIBODIES IN PLANTS" TRENDS IN BIOTECHNOLOGY, vol. 9, no. 4, 1 April 1991, pages 107-109, XP000207981 page 107, last paragraph, page 108, left hand column	1-3,6,8, 10, 21-24,29
	-/--	

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents :

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

- "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- "X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- "&" document member of the same patent family

Date of the actual completion of the international search

14 August 1998

Date of mailing of the international search report

25/08/1998

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Holtorf, S

INTERNATIONAL SEARCH REPORT

International Application No

PCT/EP 98/01731

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>HIATT A ET AL: "ASSEMBLY OF MULTIMERIC PROTEINS IN PLANT CELLS: CHARACTERISTICS AND USES OF PLANT-DERIVED ANTIBODIES" BOOKS IN SOILS PLANTS AND THE ENVIRONMENT: TRANSGENIC PLANTS: FUNDAMENTALS AND APPLICATIONS, 1 January 1993, pages 221-237, XP000569962 page 230, line 12-34</p> <p style="text-align: center;">---</p>	1-31
A	<p>EP 0 520 962 A (ENEA ENTE NUOVE TEC ;CONSIGLIO NAZIONALE RICERCHE (IT)) 30 December 1992 column 1, line 21-36; column 3, line 55 column 4, line 2; abstract, Claim 15</p> <p style="text-align: center;">---</p>	1-31
A	<p>BELL, C.W., ET AL.: "sequences of the cDNA encoding the heavy- and light-chain Fab region of an antibody to the phenylurea herbicide diuron" GENE, vol. 165, 1995, pages 323-324, XP002074415 see the whole document</p> <p style="text-align: center;">---</p>	1-31
A	<p>EP 0 716 808 A (BASF AG) 19 June 1996 see the whole document</p> <p style="text-align: center;">---</p>	1-31
A	<p>EP 0 284 419 A (ALLELIX INC) 28 September 1988 Abstract, page 2, claim 10</p> <p style="text-align: center;">---</p>	1-31
A	<p>DE 195 46 751 A (BAYER AG) 27 June 1996 Abstract, pages 4-8</p> <p style="text-align: center;">---</p>	1-31
A	<p>WO 97 04088 A (SUMITOMO CHEMICAL CO ;UNIV DUKE (US); SATO RYO (JP); BOYNTON JOHN) 6 February 1997 page 2; page 7, line 25; page 12; Example 5,7</p> <p style="text-align: center;">---</p>	1-31
A	<p>FIEDLER, U. AND CONRAD, U.: "High-level production and long-term storage of engineered antibodies in transgenic tobacco seeds" BIOTECHNOLOGY, vol. 13, October 1995, pages 1090-1093, XP002074416 Abstract, page 1090, right-hand column</p> <p style="text-align: center;">---</p> <p style="text-align: center;">-/--</p>	1-31

INTERNATIONAL SEARCH REPORT

Int. l. Application No

PCT/EP 98/01731

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ARTSAENKO, O., ET AL. : "expression of a single-chain Fv antibody against abscisic acid creates a wilted phenotype in transgenic tobacco" THE PLANT JOURNAL , vol. 8, no. 5, 1995, pages 745-750, XP002074417 seite 745; Figure. 1	1-31
A	SCHOUTEN A ET AL: "THE C-TERMINAL KDEL SEQUENCE INCREASES THE EXPRESSION LEVEL OF A SINGLE-CHAIN ANTIBODY DESIGNED TO BE TARGETED TO BOTH THE CYTOSOL AND THE SECRETORY PATHWAY IN TRANSGENIC TOBACCO" PLANT MOLECULAR BIOLOGY, vol. 30, 1996, pages 781-793, XP000677225 cited in the application	1-31

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/EP 98/01731

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0520962 A	30-12-1992	IT 1248361 B	05-01-1995
EP 0716808 A	19-06-1996	DE 4444708 A	20-06-1996
		CA 2165036 A	16-06-1996
		JP 8231309 A	10-09-1996
		US 5670454 A	23-09-1997
EP 0284419 A	28-09-1988	AU 619739 B	06-02-1992
		AU 1381388 A	29-09-1988
		AU 654686 B	17-11-1994
		AU 8797691 A	13-02-1992
		CA 1335412 A	02-05-1995
		DE 3889488 D	16-06-1994
		DE 3889488 T	15-09-1994
		DK 164888 A	28-09-1988
		ES 2052707 T	16-07-1994
		JP 1020042 A	24-01-1989
DE 19546751 A	27-06-1996	BR 9506064 A	23-12-1997
		CA 2165743 A	24-06-1996
		FR 2728433 A	28-06-1996
		IT MI952656 A	24-06-1996
		ZA 9510896 A	15-02-1996
WO 9704088 A	06-02-1997	EP 0839193 A	06-05-1998
		WO 9704089 A	06-02-1997

INTERNATIONALER RECHERCHENBERICHT

In nationales Aktenzeichen

PCT/EP 98/01731

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES

IPK 6 C12N15/82 C07K16/44 C12N15/13 A01H5/00 A01N63/00

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchiertes Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)

IPK 6 C12N C07K A01H A01N

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	ASBOTH, B.: "catalytically active antibodies for herbicide-resistant crop plant production ;antibody design for glyphosate inactivation; catalytic antibody; monoclonal antibody; abzyme may be used in the construction of a plant with glyphosate herbicide resistance (conference paper)" PLANT BIOTECHNOLGY, EC-HUNGARY WORKSHOP, 1991, Seite 21 XP002074579 siehe das ganze Dokument ---	1,3,6,8, 10,13, 21-24,29
X	SWAIN W F: "ANTIBODIES IN PLANTS" TRENDS IN BIOTECHNOLOGY, Bd. 9, Nr. 4, 1. April 1991, Seiten 107-109, XP000207981 Seite 107, letzter Absatz, Seite 108, linke Spalte --- -/--	1-3,6,8, 10, 21-24,29



Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen



Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :

"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist

"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist

"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)

"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht

"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist

"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist

"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden

"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist

"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche

14. August 1998

Absenddatum des internationalen Recherchenberichts

25/08/1998

Name und Postanschrift der Internationalen Recherchenbehörde

Europäisches Patentamt, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Bevollmächtigter Bediensteter

Holtorf, S

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN		
Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	<p>HIATT A ET AL: "ASSEMBLY OF MULTIMERIC PROTEINS IN PLANT CELLS: CHARACTERISTICS AND USES OF PLANT-DERIVED ANTIBODIES" BOOKS IN SOILS PLANTS AND THE ENVIRONMENT: TRANSGENIC PLANTS: FUNDAMENTALS AND APPLICATIONS, 1. Januar 1993, Seiten 221-237, XP000569962 Seite 230, Zeile 12-34 ---</p>	1-31
A	<p>EP 0 520 962 A (ENEA ENTE NUOVE TEC ;CONSIGLIO NAZIONALE RICERCHE (IT)) 30. Dezember 1992 Spalte 1, Zeile 21-36; Spalte 3, Zeile 55 - Spalte 4, Zeile 2; Zusammenfassung, Anspruch 15 ---</p>	1-31
A	<p>BELL, C.W., ET AL. : "sequences of the cDNA encoding the heavy- and light-chain Fab region of an antibody to the phenylurea herbicide diuron" GENE, Bd. 165, 1995, Seiten 323-324, XP002074415 siehe das ganze Dokument ---</p>	1-31
A	<p>EP 0 716 808 A (BASF AG) 19. Juni 1996 siehe das ganze Dokument ---</p>	1-31
A	<p>EP 0 284 419 A (ALLELIX INC) 28. September 1988 Zusammenfassung , Seite 2, Anspruch 10 ---</p>	1-31
A	<p>DE 195 46 751 A (BAYER AG) 27. Juni 1996 Zusammenfassung, Seite 4-8 ---</p>	1-31
A	<p>WO 97 04088 A (SUMITOMO CHEMICAL CO ;UNIV DUKE (US); SATO RYO (JP); BOYNTON JOHN) 6. Februar 1997 Seite 2; Seite 7, Zeile 25; Seite 12; Beispiele 5,7 ---</p>	1-31
A	<p>FIEDLER, U. AND CONRAD, U.: "High-level production and long-term storage of engineered antibodies in transgenic tobacco seeds" BIOTECHNOLOGY, Bd. 13, Oktober 1995, Seiten 1090-1093, XP002074416 Zusammenfassung, Seite 1090, rechte Spalte ---</p>	1-31
	-/--	

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	ARTSAENKO, O., ET AL. : "expression of a single-chain Fv antibody against abscisic acid creates a wilted phenotype in transgenic tobacco" THE PLANT JOURNAL , Bd. 8, Nr. 5, 1995, Seiten 745-750, XP002074417 Seite 745; Abb. 1	1-31
A	SCHOUTEN A ET AL: "THE C-TERMINAL KDEL SEQUENCE INCREASES THE EXPRESSION LEVEL OF A SINGLE-CHAIN ANTIBODY DESIGNED TO BE TARGETED TO BOTH THE CYTOSOL AND THE SECRETORY PATHWAY IN TRANSGENIC TOBACCO" PLANT MOLECULAR BIOLOGY, Bd. 30, 1996, Seiten 781-793, XP000677225 in der Anmeldung erwähnt	1-31

INTERNATIONALER RECHERCHENBERICHT

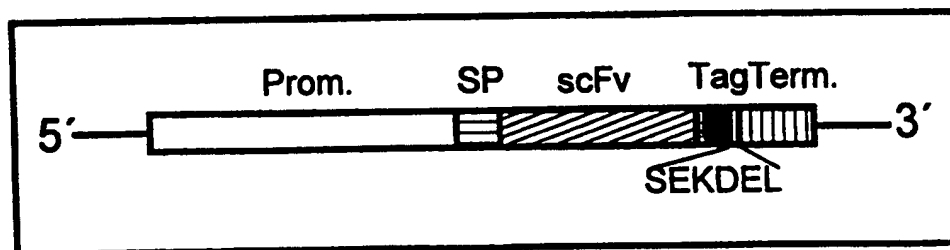
Angaben zu Veröffentlichungen, die zur selben Patentfamilie gehören

Internationales Aktenzeichen

PCT/EP 98/01731

Im Recherchenbericht angeführtes Patentdokument	Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
EP 0520962 A	30-12-1992	IT 1248361 B	05-01-1995
EP 0716808 A	19-06-1996	DE 4444708 A	20-06-1996
		CA 2165036 A	16-06-1996
		JP 8231309 A	10-09-1996
		US 5670454 A	23-09-1997
EP 0284419 A	28-09-1988	AU 619739 B	06-02-1992
		AU 1381388 A	29-09-1988
		AU 654686 B	17-11-1994
		AU 8797691 A	13-02-1992
		CA 1335412 A	02-05-1995
		DE 3889488 D	16-06-1994
		DE 3889488 T	15-09-1994
		DK 164888 A	28-09-1988
		ES 2052707 T	16-07-1994
		JP 1020042 A	24-01-1989
DE 19546751 A	27-06-1996	BR 9506064 A	23-12-1997
		CA 2165743 A	24-06-1996
		FR 2728433 A	28-06-1996
		IT MI952656 A	24-06-1996
		ZA 9510896 A	15-02-1996
WO 9704088 A	06-02-1997	EP 0839193 A	06-05-1998
		WO 9704089 A	06-02-1997

Formblatt PCT/ISA/210 (Anhang Patentfamilie)(Juli 1992)



□ USP-Promoter

▨ LeB4-Signalpeptid

■ c-myc-Tag

▤ CaMV 35 Terminator

▧ single chain Fv